5SSD0

Bayesian Machine Learning and Information Processing

Lecture Notes

Bert de Vries (Flux 7.101)

Wouter Kouw (Flux 7.060)

Magnus T. Koudahl (Flux 7.060)

5SSD0 Course Syllabus

Learning Goals

- This course provides an introduction to Bayesian machine learning and information processing systems. The Bayesian approach affords a unified and consistent treatment of many useful information processing systems.
- Upon successful completion of the course, students should be able to:
  - understand the essence of the Bayesian approach to information processing.
  - specify a solution to an information processing problem as a Bayesian inference task on a probabilistic model.
  - design a probabilistic model by a specifying a likelihood function and prior distribution;
  - Code the solution in a probabilistic programming package.
  - execute the Bayesian inference task either analytically or approximately.
  - evaluate the resulting solution by examination of Bayesian evidence.
  - be aware of the properties of commonly used probability distribitions such as the Gaussian, Gamma and multinomial distribution; models such as hidden Markov models and Gaussian mixture models; and inference methods such as the Laplace approximation, variational Bayes and message passing in a factor graph.

Entrance Requirements (pre-knowledge)

- Undergraduate courses in Linear Algebra and Probability Theory (or Statistics).
- Some scientific programming experience, eg in MATLAB or Python. In this class, we use the Julia programming language, which has a similar syntax to MATLAB, but is (close to) as fast as C.

Instructors

- Bert de Vries, rm. FLUX-7.101, overall responsible lecturer
- Wouter Kouw, rm. FLUX-7.060, responsible for Probabilistic Programming lessons
- Magnus Koudahl, rm. FLUX-7.060, teaching assistant

Important Links

- Please bookmark the following three websites:

  1. The course homepage http://bmlip.nl (or try https://biaslab.github.io/teaching/bmlip) contains links to all materials such as lecture notes and video lectures.
  2. The Piazza course site will be used for Q&A and communication.
  3. The Canvas course site will be sparingly used for communication (mostly by ESA staff)

Materials

- All materials can be accessed from the course homepage.
- Materials consist of the following resources:
  - Lecture notes (mandatory)
  - Probabilistic Programming (PP) notes (mandatory)
  - optional materials to help understand the lecture and PP notes
    - video guides to the lecture notes
    - exercises
    - (Recording of) live sessions at regular class hours for this term
    - Q&A at Piazza
    - practice exams
- Source materials are available at github repo at https://github.com/bertdv/BMLIP. You do not need to bother with this site. If you spot an error in the materials, please raise the issue at Piazza.

Lecture and PP notes and Video Guides

- The lecture and PP notes contain the mandatory materials. Some lecture notes are extended by a reading assignment, see the first cell in the lecture notes. These reading assignment are also part of the mandatory materials.
- Slides that are not required for the exam are moved to the end of the notes and preceded by an OPTIONAL SLIDES header.

- The accompanying Video Guides aim to cover just the main points or (expected) sticky issues in a lecture.

Study Guide

- Here's our recommendation on how to study for this class. Before each lecture:
  - First watch the **video guide** for that lecture
  - Then study the **lecture notes**
  - Then try to make the **exercises** for that class.
    - When you do the exercises, feel free to make use of Sam Roweis' cheat sheets for Matrix identities and Gaussian identities. Also accessible from course homepage.
  - Optionally, come to the **live class** at regular class hours to discuss remaining issues in person. The live class sessions will be recorded and posted at the class homepage.
- Aside from the regular classes, you are encouraged to disscuss any issues in Piazza. Your questions will be answered at the Piazza site by fellow students and accorded (or dorrected, amended) by the teaching staff.

Piazza (Q&A)

- We will be using Piazza for class discussion. The system is highly catered to getting you help fast and efficiently from both classmates and the teaching staff.
- Sign up for Piazza today if you have not done so. And install the Piazza app on your phone!
- The quicker you begin asking questions on Piazza (rather than via emails), the quicker you'll benefit from the collective knowledge of your classmates and instructors. We encourage you to ask questions when you're struggling to understand a concept—you can even do so anonymously.
- We will also disseminate news and announcements via Piazza.
- Unless it is a personal issue, pose your course-related questions at Piazza (in the right folder).
- Please contribute to the class by answering questions at Piazza.
  - If so desired, you can contribute anonymously.
  - Answering technical questions at Piazza is a great way to learn. If you really want to understand a topic, you should try to explain it to others.
  - Every question has just a single students' answer that students can edit collectively (and a single instructors' answer for instructors).
- You can use LaTeX in Piazza for math (and please do so!).
- Piazza has a great `search` feature. Use search before putting in new questions.

Live Class Sessions

- We will hold live class sessions at the regular class hours.
- The live classes are an opportunity to speak with the teaching staff directly about any issues.
- The live sessions will probably be short: as much as possible, I'd like to address technical questions and issues through Piazza so they are more easily accessible and searchable afterwards.

Exam Guide

- There will be a written exam in multiple-choice format.
- You are not allowed to use books nor bring printed or handwritten formula sheets to the exam. Difficult-to-remember formulas are supplied at the exam sheet.
- No smartphones at the exam.
- The tested material consists of the lecture + PP notes (+ reading assignments as assigned in the first cell/slide of each lecture notebook).
- The class homepage contains two representative practice exams from previous terms.

OPTIONAL SLIDES

Machine Learning Overview

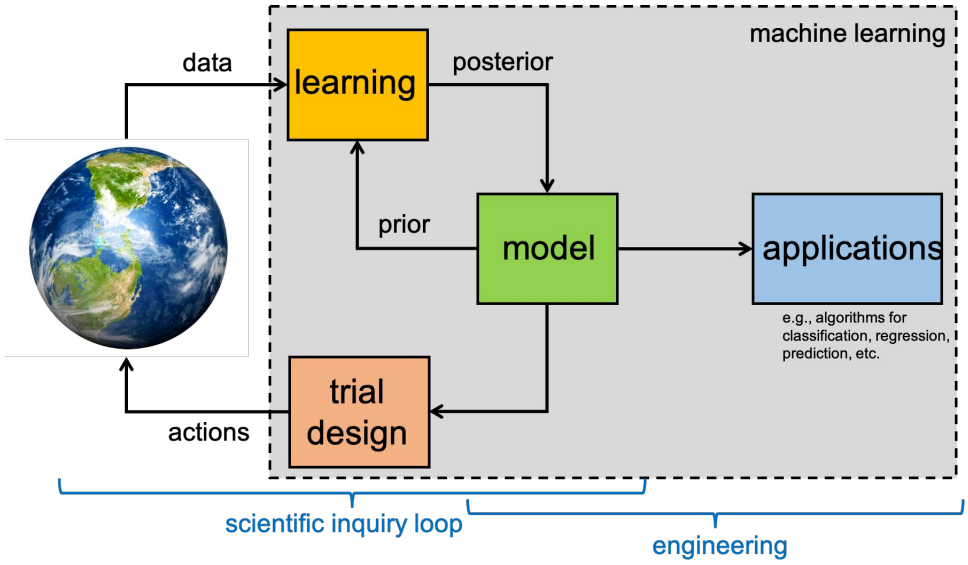
Preliminaries

- Goal
  - Top-level overview of machine learning
- Materials
  - Mandatory
    - this notebook
  - Optional
    - pre-recorded video guide and live class (2020)
    - Study Bishop pp. 1-4

What is Machine Learning?

- Machine Learning relates to **building models from data and using these models in applications**.
- **Problem**: Suppose we want to develop an algorithm for a complex process about which we have little knowledge (so hand-programming is not possible).
- **Solution**: Get the computer to develop the algorithm by itself by showing it examples of the behavior that we want.
- Practically, we choose a library of models, and write a program that picks a model and tunes it to fit the data.
- This field is known in various scientific communities with slight variations under different names such as machine learning, statistical inference, system identification, data mining, source coding, data compression, data science, etc.

Machine Learning and the Scientific Inquiry Loop

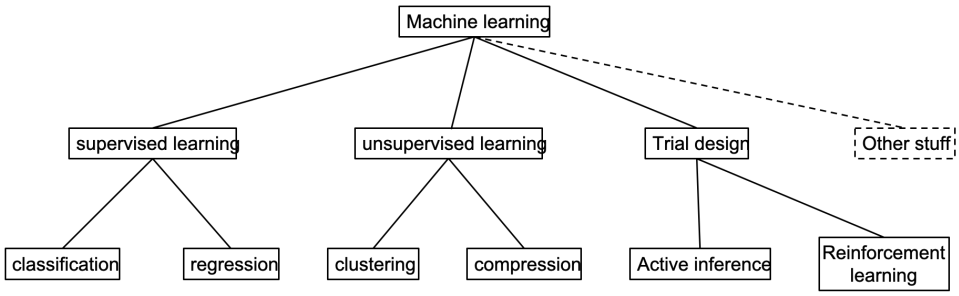


- Machine learning technology uses the scientific inquiry loop to develop models and use these models in applications.

Machine Learning is Difficult

- Modeling (Learning) Problems
  - Is there any regularity in the data anyway?
  - What is our prior knowledge and how to express it mathematically?
  - How to pick the model library?
  - How to tune the models to the data?
  - How to measure the generalization performance?
- Quality of Observed Data
  - Not enough data
  - Too much data?
  - Available data may be messy (measurement noise, missing data points, outliers)

A Machine Learning Taxonomy



- **Supervised Learning**: Given examples of inputs and corresponding desired outputs, predict outputs on future inputs.
  - Examples: classification, regression, time series prediction

- **Unsupervised Learning**: (a.k.a. **density estimation**). Given only inputs, automatically discover representations, features, structure, etc.
  - Examples: clustering, outlier detection, compression

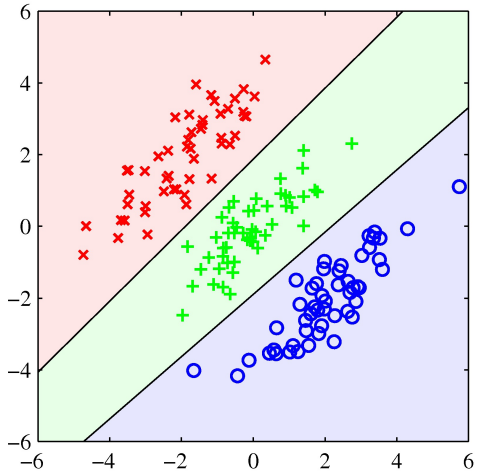
- **Trial Design**: (a.k.a. experimental design, active learning). Learn to make actions that optimize some performance criterion about the expected future.
  - Examples: playing games like chess, self-driving cars, robotics.
  - Two major approaches include **reinforcement learning** and **active inference**
    - **Reinforcement Learning**: Given an observed sequence of input signals and (occasionally observed) rewards for those inputs, *learn* to select actions that maximize *expected* future rewards.
    - **Active inference**: Given an observed sequence of input signals and a prior probability distribution about future observations, *learn* to select actions that minimize *expected* prediction errors (i.e., minimize actual minus predicted sensation).

- Other stuff, like **preference learning**, **learning to rank**, etc., can often be (re-)formulated as special cases of either a supervised, unsupervised or trial design problem.

Supervised Learning

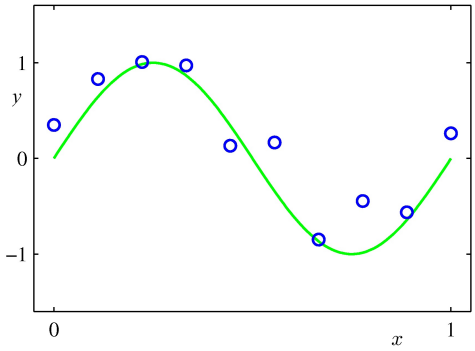
- Given observations of desired input-output behavior $D = \{(x_1, y_1), \ldots, (x_N, y_N)\}$ (with x_i inputs and y_i outputs), the goal is to estimate the conditional distribution $p(y|x)$ (i.e., how does y depend on x?).

Classification



- The target variable y is a *discrete-valued* vector representing class labels
- The special case $y \in \{\text{true}, \text{false}\}$ is called **detection**.

Regression

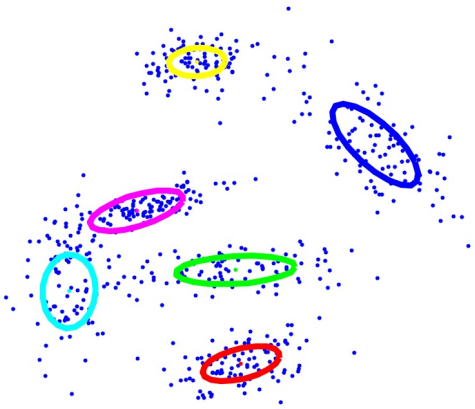


- Same problem statement as classification but now the target variable is a *real-valued* vector.
- Regression is sometimes called **curve fitting**.

Unsupervised Learning

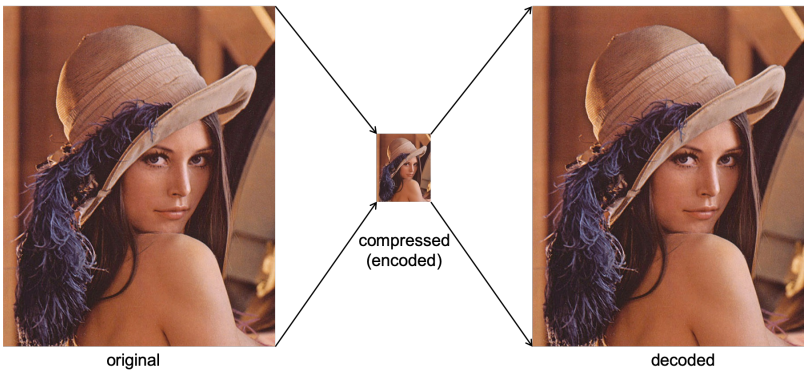
Given data $D = \{x_1, \ldots, x_N\}$, model the (unconditional) probability distribution $p(x)$ (a.k.a. **density estimation**). The two primary applications are **clustering** and **compression** (a.k.a. dimensionality reduction).

Clustering



- Group data into clusters such that all data points in a cluster have similar properties.
- Clustering can be interpreted as ''unsupervised classification''.

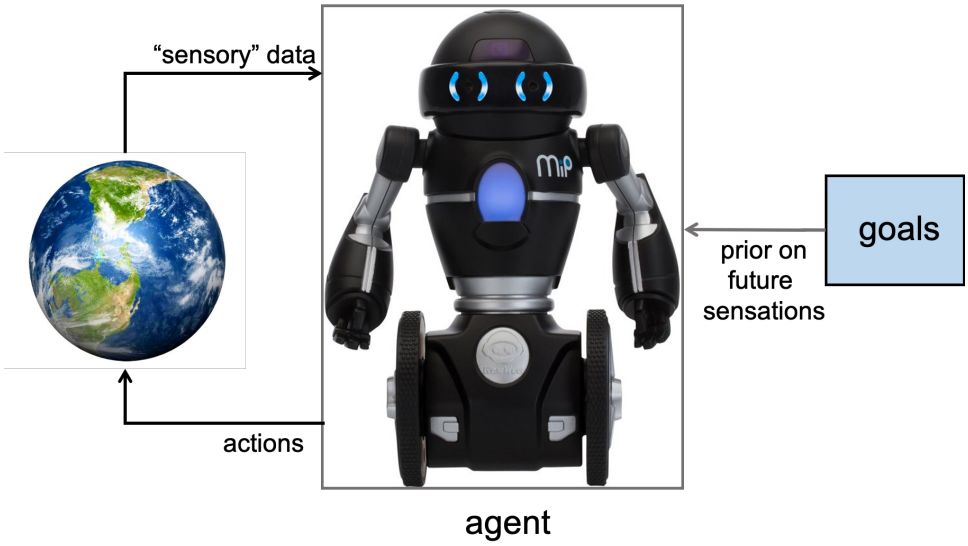
Compression / dimensionality reduction



- Output from coder is much smaller in size than original, but if coded signal if further processed by a decoder, then the result is very close (or exactly equal) to the original.
- Usually, the compressed image comprises continuously valued variables. In that case, compression can be interpreted as ''unsupervised regression''.

Trial Design

Given the state of the world (obtained from sensory data), the agent must *learn* to produce actions that optimize some performance criterion about expected future.



- In this course, we focus on the active inference approach; see Intelligent Agent lesson.

Probability Theory Review

Preliminaries

- Goal
  - Review of probability theory as a theory for rational/logical reasoning with uncertainties (i.e., a Bayesian interpretation)
- Materials
  - Mandatory
    - These lecture notes
    - Ariel Caticha - 2012 - Entropic Inference and the Foundations of Physics, pp.7-26 (sections 2.1 through 2.5), on deriving probability theory. You may skip section 2.3.4: Cox's proof (pp.15-18).
      - The assignment is only meant to appreciate how this line of "axiomatic derivation" of the rules of PT goes. I will not ask questions about any details of the derivations at the exam.
  - Optional
    - the pre-recorded video guide and live class of 2020
    - Ariel Caticha - 2012 - Entropic Inference and the Foundations of Physics, pp.7-56 (ch.2: probability)
      - Great introduction to probability theory, in particular w.r.t. its correct interpretation as a state-of-knowledge.
      - Absolutely worth your time to read the whole chapter!
    - Edwin Jaynes - 2003 - Probability Theory -- The Logic of Science.
      - Brilliant book on Bayesian view on probability theory. Just for fun, scan the annotated bibliography and references.
    - D.S. Sivia, with J. Skilling - 2006 - Data Analysis: A Bayesian Tutorial
      - Very nice short introduction to Bayesian data analysis.
    - Bishop pp. 12-24

Data Analysis: A Bayesian Tutorial

- The following is an excerpt from the book Data Analysis: A Bayesian Tutorial

Preface

As an undergraduate, I always found the subject of statistics to be rather mysterious. This topic wasn't entirely new to me, as we had been taught a little bit about probability earlier at high school; for example, I was already familiar with the binomial, Poisson and normal distributions. Most of this made sense, but only seemed to relate to things like rolling dice, flipping coins, shuffling cards and so on. However, having aspirations of becoming a scientist, what I really wanted to know was how to analyse experimental data. Thus, I eagerly looked forward to the lectures on statistics. Sadly, they were a great disappointment. Although many of the tests and procedures expounded were intuitively reasonable, there was something deeply unsatisfactory about the whole affair: there didn't seem to be any underlying basic principles! Hence, the course on 'probability and statistics' had led to an unfortunate dichotomy: probability made sense, but was just a game; statistics was important, but it was a bewildering collection of tests with little obvious rhyme or reason. While not happy with this situation, I decided to put aside the subject and concentrate on real science. After all, the predicament was just a reflection of my own inadequacies and I'd just have to work at it when the time came to really analyse my data.

The story above is not just my own, but is the all too common experience of many scientists. Fortunately, it doesn't have to be like this. What we were not told in our undergraduate lectures is that there is an alternative approach to the whole subject of data analysis which uses only probability theory. In one sense, it makes the topic of statistics entirely superfluous. In another, it provides the logical justification for many of the prevalent statistical tests and procedures, making explicit the conditions and approximations implicitly assumed in their use.

- Does this fragment resonate with your own experience?
- In this lesson we introduce *Probability Theory* (PT) again. As we will see in the next lessons, PT is all you need to make sense of machine learning, artificial intelligence, statistics, etc.

Example Problem: Disease Diagnosis

- **Problem**: Given a disease with prevalence of 1% and a test procedure with sensitivity ('true positive' rate) of 95% and specificity ('true negative' rate) of 85% , what is the chance that somebody who tests positive actually has the disease?
- **Solution**: Use probabilistic inference, to be discussed in this lecture.

The Design of Probability Theory

- Define an **event** (or "proposition") A as a statement, whose truth can be contemplated by a person, e.g.,

$$A = \text{'there is life on Mars'}$$

- If we assume the fact

$$I = \text{'All known life forms require water'}$$

  and a new piece of information

$$x = \text{'There is water on Mars'}$$

  becomes available, how *should* our degree of belief in event A be affected (if we were rational)?

- Richard T. Cox, 1946 developed a **calculus for rational reasoning** about how to represent and update the degree of *beliefs* about the truth value of events when faced with new information.
- In developing this calculus, only some very agreeable assumptions were made, e.g.,
  - (Transitivity). If the belief in A is greater than the belief in B, and the belief in B is greater than the belief in C, then the belief in A must be greater than the belief in C.
  - (Consistency). If the belief in an event can be inferred in two different ways, then the two ways must agree on the resulting belief.
- This effort resulted in confirming that the sum and product rules of Probability Theory are the **only** proper rational way to process belief intensities.
- $\Rightarrow$ Probability theory (PT) provides *the* **theory of optimal processing of incomplete information** (see Cox theorem, and Caticha, pp.7-24), and as such provides the (only) proper quantitative framework for drawing conclusions from a finite (read: incomplete) data set.

Why Probability Theory for Machine Learning?

- Machine learning concerns drawing conclusions about model parameter settings from (a finite set of) data and therefore PT provides the *optimal calculus for machine learning*.

- In general, nearly all interesting questions in machine learning can be stated in the following form (a conditional probability):

$$p(\texttt{whatever-we-want-to-know} \,|\, \texttt{whatever-we-do-know})$$

where $p(a|b)$ means the probability that a is true, given that b is true.

- Examples
  - Predictions

$$p(\texttt{future-observations} \,|\, \texttt{past-observations})$$

  - Classify a received data point x

$$p(\, x\texttt{-belongs-to-class-}k \,|\, x\,)$$

  - Update a model based on a new observation

$$p(\texttt{model-parameters} \,|\, \texttt{new-observation}, \texttt{past-observations})$$

Frequentist vs. Bayesian Interpretation of Probabilities

- The interpretation of a probability as a **degree-of-belief** about the truth value of an event is also called the **Bayesian** interpretation.

- In the **Bayesian** interpretation, the probability is associated with a **state-of-knowledge** (usually held by a person).
  - For instance, in a coin tossing experiment, $p(\texttt{tail}) = 0.4$ should be interpreted as the belief that there is a 40% chance that `tail` comes up if the coin were tossed.
  - Under the Bayesian interpretation, PT calculus (sum and product rules) **extends boolean logic to rational reasoning with uncertainty**.

- The Bayesian interpretation contrasts with the **frequentist** interpretation of a probability as the relative frequency that an event would occur under repeated execution of an experiment.

  - For instance, if the experiment is tossing a coin, then $p(\texttt{tail}) = 0.4$ means that in the limit of a large number of coin tosses, 40% of outcomes turn up as `tail`.

- The Bayesian viewpoint is more generally applicable than the frequentist viewpoint, e.g., it is hard to apply the frequentist viewpoint to events like '`it will rain tomorrow`'.

- The Bayesian viewpoint is clearly favored in the machine learning community. (In this class, we also strongly favor the Bayesian interpretation).

Probability Theory Notation

events

- Define an **event** A as a statement, whose truth can be contemplated by a person, e.g.,

$$A = \text{'it will rain tomorrow'}$$

- We write the denial of A, i.e. the event **not**-A, as $\bar{A}$.

- Given two events A and B, we write the **conjunction** "$A \wedge B$" as "A, B" or "AB". The conjunction AB is true only if both A and B are true.

- We will write the **disjunction** "$A \vee B$" as "$A + B$", which is true if either A or B is true or both A and B are true.

- Note that, if X is a variable, then an assignment $X = x$ (with x a value, e.g., $X = 5$) can be interpreted as an event.

probabilities

- For any event A, with background knowledge I, the **conditional probability of** A **given** I, is written as

$$p(A|I)\,.$$

- All probabilities are in principle conditional probabilities of the type $p(A|I)$, since there is always some background knowledge.

Unfortunately, PT notation is usually rather sloppy :(

- We often write $p(A)$ rather than $p(A|I)$ if the background knowledge I is assumed to be obviously present. E.g., $p(A)$ rather than $p(A|\text{the-sun-comes-up-tomorrow})$.
- (In the context of random variable assignments) we often write $p(x)$ rather than $p(X=x)$, assuming that the reader understands the context.
- In an apparent effort to further abuse notational conventions, $p(X)$ denotes the full distribution over random variable X, i.e., the distribution for all assignments for X.
- If X is a *discretely* valued variable, then $p(X=x)$ is a probability *mass* function (PMF) with $0 \leq p(X=x) \leq 1$ and normalization $\sum_x p(x) = 1$.
- If X is *continuously* valued, then $p(X=x)$ is a probability *density* function (PDF) with $p(X=x) \geq 0$ and normalization $\int_x p(x)\mathrm{d}x = 1$.
  - Note that if X is continuously valued, then the value of the PDF $p(x)$ is not necessarily ≤ 1. E.g., a uniform distribution on the continuous domain $[0, .5]$ has value $p(x) = 2$.
- Often, we do not bother to specify if $p(x)$ refers to a continuous or discrete variable.

Probability Theory Calculus

- Let $p(A|I)$ indicate the belief in event A, given that I is true.
- The following product and sum rules are also known as the **axioms of probability theory**, but as discussed above, under some mild assumptions, they can be derived as the unique rules for *rational reasoning under uncertainty* (Cox theorem, 1946, and Caticha, 2012, pp.7-26).
- **Sum rule**. The disjunction for two events A and B given background I is given by

$$\boxed{p(A+B|I) = p(A|I) + p(B|I) - p(A, B|I)}$$

- **Product rule**. The conjuction of two events A and B with given background I is given by

$$\boxed{p(A, B|I) = p(A|B, I)\, p(B|I)}$$

- **All legitimate probabilistic relations can be derived from the sum and product rules!**

Independent and Mutually Exclusive Events

- Two events A and B are said to be **independent** if the probability of one is not altered by information about the truth of the other, i.e., $p(A|B) = p(A)$
  - $\Rightarrow$ If A and B are independent, given I, then the product rule simplifies to

$$p(A, B|I) = p(A|I)p(B|I)$$

- Two events A_1 and A_2 are said to be **mutually exclusive** if they cannot be true simultanously, i.e., if $p(A_1, A_2) = 0$.
  - $\Rightarrow$ For mutually exclusive events, the sum rule simplifies to

$$p(A_1 + A_2) = p(A_1) + p(A_2)$$

- A set of events $A_1, A_2, \ldots, A_N$ is said to be **collectively exhaustive** if one of the statements is necessarily true, i.e., $A_1 + A_2 + \cdots + A_N = \text{TRUE}$, or equivalently

$$p(A_1 + A_2 + \cdots + A_N) = 1$$

- Note that, if $A_1, A_2, \ldots, A_n$ are both **mutually exclusive** and **collectively exhausitive** (MECE) events, then

$$\sum_{n=1}^{N} p(A_n) = p(A_1 + \ldots + A_N) = 1$$

  - More generally, if $\{A_n\}$ are MECE events, then $\sum_{n=1}^{N} p(A_n, B) = p(B)$

The Sum Rule and Marginalization

- We mentioned that every inference problem in PT can be evaluated through the sum and product rules. Next, we present two useful corollaries: (1) *Marginalization* and (2) *Bayes rule*
- If $X \in \mathcal{X}$ and $Y \in \mathcal{Y}$ are random variables over finite domains, than it follows from the above considerations about MECE events that

$$\sum_{Y \in \mathcal{Y}} p(X, Y) = p(X)\,.$$

- Summing Y out of a joint distribution $p(X,Y)$ is called **marginalization** and the result $p(X)$ is sometimes referred to as the **marginal probability**.
- Note that this is just a **generalized sum rule**. In fact, Bishop (p.14) (and some other authors as well) calls this the sum rule.
- Of course, in the continuous domain, the (generalized) sum rule becomes

$$p(X) = \int p(X,Y)\,\mathrm{d}Y$$

The Product Rule and Bayes Rule

- Consider two variables D and θ; it follows from symmetry arguments that

$$p(D,\theta) = p(D|\theta)p(\theta) = p(\theta|D)p(D)$$

  and hence that

$$p(\theta|D) = \frac{p(D|\theta)}{p(D)}p(\theta)\,.$$

- This formula is called **Bayes rule** (or Bayes theorem). While Bayes rule is always true, a particularly useful application occurs when D refers to an observed data set and θ is set of model parameters. In that case,
  - the **prior** probability $p(\theta)$ represents our **state-of-knowledge** about proper values for θ, before seeing the data D.
  - the **posterior** probability $p(\theta|D)$ represents our state-of-knowledge about θ after we have seen the data.

⇒ Bayes rule tells us how to update our knowledge about model parameters when facing new data. Hence,

Bayes rule is the fundamental rule for learning from data!

Bayes Rule Nomenclature

- Some nomenclature associated with Bayes rule:

$$\underbrace{p(\theta|D)}_{\text{posterior}} = \frac{\overbrace{p(D|\theta)}^{\text{likelihood}} \times \overbrace{p(\theta)}^{\text{prior}}}{\underbrace{p(D)}_{\text{evidence}}}$$

- Note that the evidence (a.k.a. *marginal likelihood*) can be computed from the numerator through marginalization since

$$p(D) = \int p(D,\theta)\,\mathrm{d}\theta = \int p(D|\theta)\,p(\theta)\,\mathrm{d}\theta$$

- Hence, having access to likelihood and prior is in principle sufficient to compute both the evidence and the posterior. To emphasize that point, Bayes rule is sometimes written as a transformation:

$$\underbrace{\underbrace{p(\theta|D)}_{\text{posterior}} \cdot \underbrace{p(D)}_{\text{evidence}}}_{\text{this is what we want to compute}} = \underbrace{\underbrace{p(D|\theta)}_{\text{likelihood}} \cdot \underbrace{p(\theta)}_{\text{prior}}}_{\text{this is available}}$$

- For a given data set D, the posterior probabilities of the parameters scale relatively against each other as

$$p(\theta|D) \propto p(D|\theta)p(\theta)$$

- ⇒ All that we can learn from the observed data is contained in the likelihood function $p(D|\theta)$. This is called the **likelihood principle**.

The Likelihood Function vs the Sampling Distribution

- Consider a distribution $p(D|\theta)$, where D relates to variables that are observed (i.e., a "data set") and θ are model parameters.
- In general, $p(D|\theta)$ is just a function of the two variables D and θ. We distinguish two interpretations of this function, depending on which variable is observed (or given by other means).
- The **sampling distribution** (a.k.a. the **data-generating** distribution)

$$p(D|\theta = \theta_0)$$

  (which is a function of D only) describes a probability distribution for data D, assuming that it is generated by the given model with parameters fixed at $\theta = \theta_0$.

- In a machine learning context, often the data is observed, and θ is the free variable. In that case, for given observations $D = D_0$, the **likelihood function** (which is a function only of the model parameters θ) is defined as

$$\mathrm{L}(\theta) \triangleq p(D = D_0|\theta)$$

- Note that $\mathrm{L}(\theta)$ is not a probability distribution for θ since in general $\sum_\theta \mathrm{L}(\theta) \neq 1$.

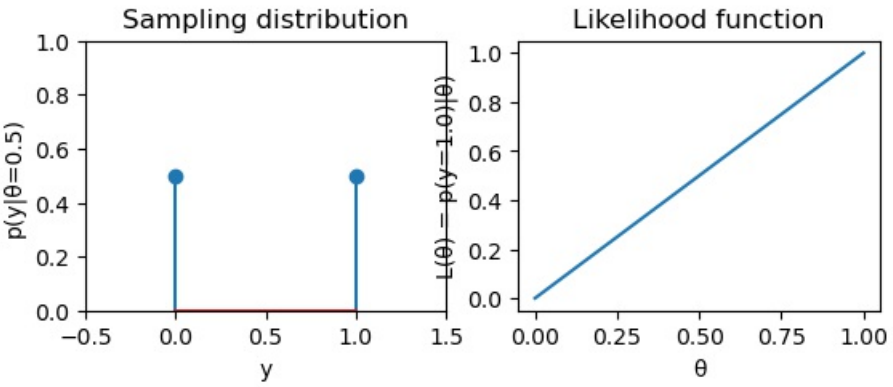
Code Example: Sampling Distribution and Likelihood Function for the Coin Toss

Consider the following simple model for the outcome (head or tail) $y \in \{0, 1\}$ of a biased coin toss with parameter $\theta \in [0, 1]$:

$$p(y|\theta) \triangleq \theta^y (1-\theta)^{1-y}$$

We can plot both the sampling distribution $p(y|\theta = 0.8)$ and the likelihood function $L(\theta) = p(y = 0|\theta)$.

In [1]:
```
using PyPlot
#using Plots
p(y,θ) = θ.^y .* (1 .- θ).^(1 .- y)
f = figure()

θ = 0.5 # Set parameter
# Plot the sampling distribution
subplot(221); stem([0,1], p([0,1],θ));
title("Sampling distribution");
xlim([-0.5,1.5]); ylim([0,1]); xlabel("y"); ylabel("p(y|θ=$(θ))");

subplot(222);
_θ = 0:0.01:1
y = 1.0 # Plot p(y=1 | θ)
plot(_θ,p(y,_θ))
title("Likelihood function");
xlabel("θ");
ylabel("L(θ) = p(y=$y)|θ)");
```



The (discrete) sampling distribution is a valid probability distribution. However, the likelihood function $L(\theta)$ clearly isn't, since $\int_0^1 L(\theta) \mathrm{d}\theta \neq 1$.

Probabilistic Inference

- **Probabilistic inference** refers to computing

$$p(\text{ whatever-we-want-to-know} | \text{ whatever-we-already-know })$$

  - For example:

$$p(\text{ Mr.S.-killed-Mrs.S. } | \text{ he-has-her-blood-on-his-shirt })$$
$$p(\text{ transmitted-codeword} | \text{ received-codeword })$$

- This can be accomplished by repeated application of sum and product rules.

- In particular, consider a joint distribution $p(X, Y, Z)$. Assume we are interested in $p(X|Z)$:

$$p(X|Z) \overset{p}{=} \frac{p(X,Z)}{p(Z)} \overset{s}{=} \frac{\sum_Y p(X,Y,Z)}{\sum_{X,Y} p(X,Y,Z)},$$

  where the 's' and 'p' above the equality sign indicate whether the sum or product rule was used.

- In the rest of this course, we'll encounter many long probabilistic derivations. For each manipulation, you should be able to associate an 's' (for sum

rule), a 'p' (for product or Bayes rule) or an 'm' (for a simplifying model assumption) above any equality sign.

Working out the example problem: Disease Diagnosis

- **Problem**: Given a disease D with prevalence of 1% and a test procedure T with sensitivity ('true positive' rate) of 95% and specificity ('true negative' rate) of 85%, what is the chance that somebody who tests positive actually has the disease?

- **Solution**: The given data are $p(D=1) = 0.01$, $p(T=1|D=1) = 0.95$ and $p(T=0|D=0) = 0.85$. Then according to Bayes rule,

$$\begin{aligned} p(D=1|T=1) \\ &\overset{p}{=} \frac{p(T=1|D=1)p(D=1)}{p(T=1)} \\ &\overset{s}{=} \frac{p(T=1|D=1)p(D=1)}{p(T=1|D=1)p(D=1)+p(T=1|D=0)p(D=0)} \\ &= \frac{0.95 \times 0.01}{0.95 \times 0.01 + 0.15 \times 0.99} = 0.0601 \end{aligned}$$

Inference Exercise: Bag Counter

- **Problem**: A bag contains one ball, known to be either white or black. A white ball is put in, the bag is shaken, and a ball is drawn out, which proves to be white. What is now the chance of drawing a white ball?

- **Solution**: Again, use Bayes and marginalization to arrive at $p(\text{white}|\text{data}) = 2/3$, see the Exercises notebook.

- $\Rightarrow$ Note that probabilities describe **a person's state of knowledge** rather than a 'property of nature'.

Inference Exercise: Causality?

- **Problem**: A dark bag contains five red balls and seven green ones. (a) What is the probability of drawing a red ball on the first draw? Balls are not returned to the bag after each draw. (b) If you know that on the second draw the ball was a green one, what is now the probability of drawing a red ball on the first draw?

- **Solution**: (a) $5/12$. (b) $5/11$, see the Exercises notebook.

- $\Rightarrow$ Again, we conclude that conditional probabilities reflect **implications for a state of knowledge** rather than temporal causality.

Moments of the PDF

- Consider a distribution $p(x)$. The **expected value** or **mean** is defined as

$$\mu_x = \mathbb{E}[x] \triangleq \int x\, p(x)\, \mathrm{d}x$$

- The **variance** of x is defined as

$$\Sigma_x \triangleq \mathbb{E}\left[(x-\mu_x)(x-\mu_x)^T\right]$$

- The **covariance** matrix between *vectors* x and y is defined as

$$\begin{aligned} \Sigma_{xy} &\triangleq \mathbb{E}\left[(x-\mu_x)(y-\mu_y)^T\right] \\ &= \mathbb{E}\left[(x-\mu_x)(y^T-\mu_y^T)\right] \\ &= \mathbb{E}[xy^T] - \mu_x\mu_y^T \end{aligned}$$

  - Clearly, if x and y are independent, then $\Sigma_{xy} = 0$, since $\mathbb{E}[xy^T] = \mathbb{E}[x]\mathbb{E}[y^T] = \mu_x\mu_y^T$.

Linear Transformations

- Consider an arbitrary distribution $p(X)$ with mean μ_x and variance Σ_x and the linear transformation

$$Z = AX + b.$$

- No matter the specification of $p(X)$, we can derive that (see Exercises notebook)

$$\begin{aligned} \mu_z &= A\mu_x + b && \text{(SRG-3a)} \\ \Sigma_z &= A\,\Sigma_x\,A^T && \text{(SRG-3b)} \end{aligned}$$

  - (The tag (SRG-3a) refers to the corresponding eqn number in Sam Roweis Gaussian identities notes.)

PDF for the Sum of Two Variables

- Given eqs SRG-3a and SRG-3b (previous cell), you should now be able to derive the following: for any distribution of variable x and y and sum $z = x + y$ (proof by Exercise)

$$\mu_z = \mu_x + \mu_y$$
$$\Sigma_z = \Sigma_x + \Sigma_y + 2\Sigma_{xy}$$

- Clearly, it follows that if x and y are **independent**, then

$$\Sigma_z = \Sigma_x + \Sigma_y$$

- More generally, given two **independent** variables x and y, with PDF's $p_x(x)$ and $p_y(y)$. The PDF $p_z(z)$for $z = x + y$ is given by the **convolution**

$$p_z(z) = \int_{-\infty}^{\infty} p_x(x) p_y(z - x)\, \mathrm{d}x$$

- **Proof**: Let $p_z(z)$ be the probability that z has value z. This occurs if x has some value x and at the same time $y = z - x$, with joint probability $p_x(x)p_y(z - x)$. Since x can be any value, we sum over all possible values for x to get $p_z(z) = \int_{-\infty}^{\infty} p_x(x) p_y(z - x)\, \mathrm{d}x$
  - Note that $p_z(z) \neq p_x(x) + p_y(y)$!!

- https://en.wikipedia.org/wiki/List_of_convolutions_of_probability_distributions shows how these convolutions work out for a few common probability distributions.

- In linear stochastic systems theory, the Fourier Transform of a PDF (i.e., the characteristic function) plays an important computational role. Why?

Code Example: Sum of Two Gaussian Distributed Variables

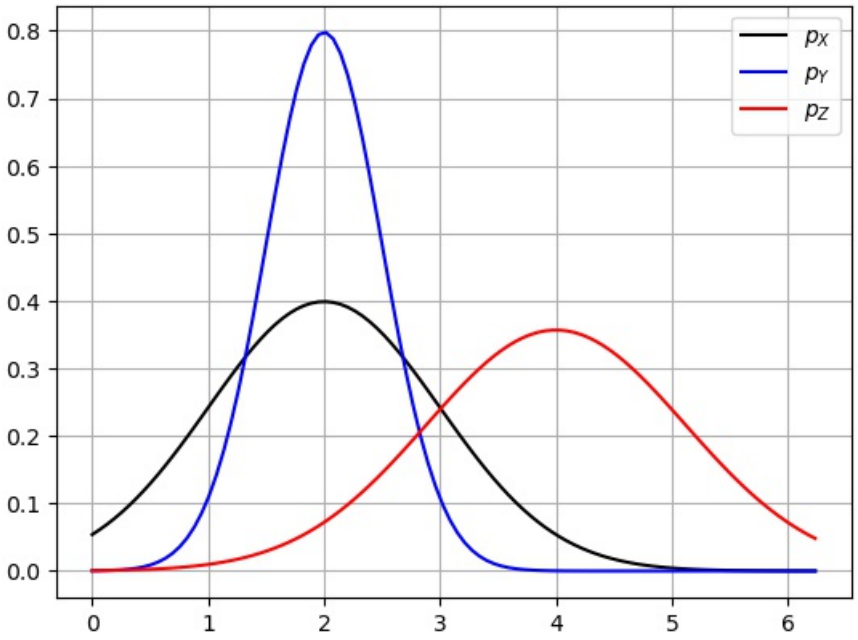
- Consider the PDF of the sum of two independent Gaussian distributed x and y:

$$p_X(x) = \mathcal{N}(x \mid \mu_X, \sigma_X^2)$$
$$p_Y(y) = \mathcal{N}(y \mid \mu_Y, \sigma_Y^2)$$

- Let $z = x + y$. Performing the convolution (nice exercise) yields a Gaussian PDF for z:

$$p_Z(z) = \mathcal{N}(z \mid \mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2).$$

In [2]:
```
using PyPlot, Distributions
μx = 2.
σx = 1.
μy = 2.
σy = 0.5
μz = μx+μy; σz = sqrt(σx^2 + σy^2)
x = Normal(μx, σx)
y = Normal(μy, σy)
z = Normal(μz, σz)
range_min = minimum([μx-2*σx, μy-2*σy, μz-2*σz])
range_max = maximum([μx+2*σx, μy+2*σy, μz+2*σz])
range_grid = range(range_min, stop=range_max, length=100)
plot(range_grid, pdf.(x,range_grid), "k-")
plot(range_grid, pdf.(y,range_grid), "b-")
plot(range_grid, pdf.(z,range_grid), "r-")
legend([L"p_X", L"p_Y", L"p_Z"])
grid()
```



PDF for the Product of Two Variables

- For two continuous random **independent** variables X and Y, with PDF's $p_x(x)$ and $p_y(y)$, the PDF of $Z = XY$ is given by

$$p_z(z) = \int_{-\infty}^{\infty} p_x(x)\, p_y(z/x)\, \frac{1}{|x|}\, \mathrm{d}x$$

- For proof, see https://en.wikipedia.org/wiki/Product_distribution

- Generally, this integral does not lead to an analytical expression for $p_z(z)$. For example, **the product of two independent variables that are both normally (Gaussian) distributed does not lead to a normal distribution**.
  - Exception: the distribution of the product of two variables that both have log-normal distributions is again a lognormal distribution.
    - (If X has a normal distribution, then $Y = \exp(X)$ has a log-normal distribution.)

Variable Transformations

- Suppose x is a **discrete** random variable with probability **mass** function $P_x(x)$, and $y = h(x)$ is a one-to-one function with $x = g(y) = h^{-1}(y)$. Then

$$P_y(y) = P_x(g(y))\,.$$

- **Proof**: $P_y(\hat{y}) = P(y = \hat{y}) = P(h(x) = \hat{y}) = P(x = g(\hat{y})) = P_x(g(\hat{y}))$. □

- If x is defined on a **continuous** domain, and $p_x(x)$ is a probability **density** function, then probability mass is represented by the area under a (density) curve. Let $a = g(c)$ and $b = g(d)$. Then

$$\begin{aligned} P(a \le x \le b) &= \int_a^b p_x(x)\mathrm{d}x \\ &= \int_{g(c)}^{g(d)} p_x(x)\mathrm{d}x \\ &= \int_c^d p_x(g(y))\mathrm{d}g(y) \\ &= \int_c^d \underbrace{p_x(g(y))g'(y)}_{p_y(y)}\mathrm{d}y \\ &= P(c \le y \le d) \end{aligned}$$

- Equating the two probability masses leads to identificaiton of the relation

$$p_y(y) = p_x(g(y))g'(y)\,,$$

  which is also known as the Change-of-Variable theorem.

- If the tranformation $y = h(x)$ is not invertible, then $x = g(y)$ does not exist. In that case, you can still work out the transformation by equating equivalent probability masses in the two domains.

Example: Transformation of a Gaussian Variable

- Let $p_x(x) = \mathcal{N}(x|\mu,\sigma^2)$ and $y = \frac{x-\mu}{\sigma}$.
- **Problem**: What is $p_y(y)$?
- **Solution**: Note that $h(x)$ is invertible with $x = g(y) = \sigma y + \mu$. The change-of-variable formula leads to

$$\begin{aligned} p_y(y) &= p_x(g(y)) \cdot g'(y) \\ &= p_x(\sigma y + \mu) \cdot \sigma \\ &= \frac{1}{\sigma\sqrt{(2\pi)}} \exp\left(-\frac{(\sigma y + \mu - \mu)^2}{2\sigma^2}\right) \cdot \sigma \\ &= \frac{1}{\sqrt{(2\pi)}} \exp\left(-\frac{y^2}{2}\right) \\ &= \mathcal{N}(y|0,1) \end{aligned}$$

Summary

- Probabilities should be interpretated as degrees of belief, i.e., a state-of-knowledge, rather than a property of nature.
- We can do everything with only the **sum rule** and the **product rule**. In practice, **Bayes rule** and **marginalization** are often very useful for inference, i.e., for computing

$$p(\text{ what-we-want-to-know} \,|\, \text{what-we-already-know}\,).$$

- Bayes rule

$$p(\theta|D) = \frac{p(D|\theta)p(\theta)}{p(D)}$$

  is the fundamental rule for learning from data!
- For a variable x with distribution $p(X)$ with mean μ_x and variance Σ_x, the mean and variance of the **Linear Transformation** $z = Ax + b$ is given by

$$\mu_z = A\mu_x + b \tag{SRG-3a}$$
$$\Sigma_z = A\,\Sigma_x\,A^T \tag{SRG-3b}$$

Bayesian Machine Learning

Preliminaries

- Goals
  - Introduction to Bayesian (i.e., probabilistic) modeling
- Materials
  - Mandatory
    - These lecture notes
  - Optional
    - Bishop pp. 68-74 (on the coin toss example)
    - Ariel Caticha - 2012 - Entropic Inference and the Foundations of Physics, pp.35-44 (section 2.9, on deriving Bayes rule for updating probabilities)

Challenge: Predicting a Coin Toss

- **Problem**: We observe a the following sequence of heads (h) and tails (t) when tossing the same coin repeatedly

$$D = \{hthhtth\}\,.$$

- What is the probability that heads comes up next?
- **Solution**: later in this lecture.

The Bayesian Machine Learning Framework

- Suppose that your application is to predict a future observation x, based on N past observations $D = \{x_1, \ldots, x_N\}$.
- The Bayesian design approach to solving this task involves four stages:

```
REPEAT   1- Model specification   2- Parameter estimation   3- Model evaluation UNTIL model performance is satisfactory
  4- Apply model
```

- Intosses principle, based on the model evaluationtosses results, you may want to re-specify your model and *repeat* the design process (a few times), until model performance is acceptable.

- Next, we discuss these four stages in a bit more detail.

(1) Model specification

- Your first task is to propose a probabilistic model (m) for generating the observations x.

- A probabilistic model m consists of a joint distribution $p(x, \theta | m)$ that relates observations x to model parameters θ. Usually, the model is proposed in the form of a data generating distribution $p(x|\theta, m)$ and a prior $p(\theta|m)$.

- *You* are responsible to choose the data generating distribution $p(x|\theta)$ based on your physical understanding of the data generating process. (For brevity, since we are working on one given model m, we drop the given dependency on m from the notation).

- *You* must also choose the prior $p(\theta)$ to reflect what you know about the parameter values before you see the data D.

(2) Parameter estimation

- Note that, for a given data set $D = \{x_1, x_2, \ldots, x_N\}$ with *independent* observations x_n, the likelihood factorizes as

$$p(D|\theta) = \prod_{n=1}^{N} p(x_n|\theta) \,,$$

  so usually you select a model for generating one observation x_n and then use (in-)dependence assumptions to combine these models into a likelihood function for the model parameters.

- The likelihood and prior both contain information about the model parameters. Next, you use Bayes rule to fuse these two information sources into a posterior distribution for the parameters:

$$\underbrace{p(\theta|D)}_{\text{posterior}} = \frac{p(D|\theta)p(\theta)}{p(D)} \propto \underbrace{p(D|\theta)}_{\text{likelihood}} \cdot \underbrace{p(\theta)}_{\text{prior}}$$

- Note that there's **no need for you to design some clever parameter estimation algorithm**. Bayes rule *is* the parameter estimation algorithm. The only complexity lies in the computational issues!

- This "recipe" works only if the right-hand side (RHS) factors can be evaluated; the computational details can be quite challenging and this is what machine learning is about.

- $\Rightarrow$ **Machine learning is EASY, apart from computational details :)**

(3) Model Evaluation

- In the framework above, parameter estimation was executed by "perfect" Bayesian reasoning. So is everything settled now?

- No, there appears to be one remaining problem: how good really were our model assumptions $p(x|\theta)$ and $p(\theta)$? We want to "score" the model performance.

- Note that this question is only interesting in practice if we have alternative models to choose from. After all, if you don't have an alternative model, any value for the model evidence would still not lead you to switch to another model.

- Let's assume that we have more candidate models, say $\mathcal{M} = \{m_1, \ldots, m_K\}$ where each model relates to specific prior $p(\theta|m_k)$ and likelihood $p(D|\theta, m_k)$? Can we evaluate the relative performance of a model against another model from the set?

- Start again with **model specification**. *You* must now specify a prior $p(m_k)$ (next to the likelihood $p(D|\theta, m_k)$ and prior $p(\theta|m_k)$) for each of the models and then solve the desired inference problem:

$$\begin{aligned}
\underbrace{p(m_k|D)}_{\substack{\text{model}\\\text{posterior}}} &= \frac{p(D|m_k)p(m_k)}{p(D)} \\
&\propto p(m_k) \cdot p(D|m_k) \\
&= p(m_k) \cdot \int_\theta p(D, \theta|m_k)\,\mathrm{d}\theta \\
&= \underbrace{p(m_k)}_{\substack{\text{model}\\\text{prior}}} \cdot \underbrace{\int_\theta \underbrace{p(D|\theta, m_k)}_{\text{likelihood}} \underbrace{p(\theta|m_k)}_{\text{prior}}\,\mathrm{d}\theta}_{\substack{\text{evidence } p(D|m_k)\\=\text{ model likelihood}}}
\end{aligned}$$

- Note that, to evaluate the posterior for a model, you must calculate the "evidence", which can be interpreted as a likelihood function for the model.
- You can now compare posterior distributions $p(m_k|D)$ for a set of models $\{m_k\}$ and decide on the merits of each model relative to alternative models. This procedure is called **Bayesian model comparison**.
- $\Rightarrow$ In a Bayesian framework, **model estimation** follows the same recipe as parameter estimation; it just works at one higher hierarchical level. Compare the required calulations:

$$\begin{aligned} p(\theta|D) &\propto p(D|\theta)p(\theta) && \text{(parameter estimation)} \\ p(m_k|D) &\propto p(D|m_k)p(m_k) && \text{(model comparison)} \end{aligned}$$

- Again, **no need to invent a special algorithm for estimating the performance of your model**. Straightforward application of probability theory takes care of all that.
- In principle, you could proceed with asking how good your choice for the candidate model set $\mathcal{M}$ was. You would have to provide a set of alternative model sets $\{\mathcal{M}_1, \mathcal{M}_2, \ldots, \mathcal{M}_M\}$ with priors $p(\mathcal{M}_m)$ for each set and compute posteriors $p(\mathcal{M}_m|D)$. And so forth ...
- With the (relative) performance evaluation scores of your model in hand, you could now re-specify your model (hopefully an improved model) and *repeat* the design process until the model performance score is acceptable.
- As an aside, in the (statistics and machine learning) literature, performance comparison between two models is often reported by the Bayes Factor, which is defined as the ratio of model evidences:

$$\begin{aligned} \underbrace{\frac{p(D|m_1)}{p(D|m_2)}}_{\text{Bayes Factor}} &= \frac{\frac{p(D,m_1)}{p(m_1)}}{\frac{p(D,m_2)}{p(m_2)}} \\ &= \frac{p(D,m_1)}{p(m_1)} \cdot \frac{p(m_2)}{p(D,m_2)} \\ &= \frac{p(m_1|D)p(D)}{p(m_1)} \cdot \frac{p(m_2)}{p(m_2|D)p(D)} \\ &= \underbrace{\frac{p(m_1|D)}{p(m_2|D)}}_{\substack{\text{posterior} \\ \text{ratio}}} \cdot \underbrace{\frac{p(m_2)}{p(m_1)}}_{\substack{\text{prior} \\ \text{ratio}}} \end{aligned}$$

  - Hence, for equal model priors ($p(m_1) = p(m_2) = 0.5$), the Bayes Factor reports the posterior probability ratio for the two models.
- In principle, any decision about which is the better model has accepted some ad hocery, but Jeffreys (1961) advises the following interpretation of the log-Bayes factor $log_{10}B_{12} = \log_{10}\frac{p(D|m_1)}{p(D|m_2)}$:

$^{\log_{10}B_{12}}$	**Evidence for** $_{m_1}$
0 to 0.5	not worth mentioning
0.5 to 1	substantial
1 to 2	strong
>2	decisive

(4) Prediction

- Once we are satisfied with the evidence for a (trained) model, we can apply the model to our prediction/classification/etc task.
- Given the data D, our knowledge about the yet unobserved datum x is captured by (everything is conditioned on the selected model)

$$\begin{aligned} p(x|D) &\overset{s}{=} \int p(x, \theta|D)\,\mathrm{d}\theta \\ &\overset{p}{=} \int p(x|\theta, D)p(\theta|D)\,\mathrm{d}\theta \\ &\overset{m}{=} \int \underbrace{p(x|\theta)}_{\text{data generation dist.}} \cdot \underbrace{p(\theta|D)}_{\text{posterior}}\,\mathrm{d}\theta \end{aligned}$$

- In the last equation, the simplification $p(x|\theta, D) = p(x|\theta)$ follows from our model specification. We assumed a *parametric* data generating distribution $p(x|\theta)$ with no explicit dependency on the data set D.
- Again, **no need to invent a special prediction algorithm**. Probability theory takes care of all that. The complexity of prediction is just computational: how to carry out the marginalization over θ.
- Note that the application of the learned posterior $p(\theta|D)$ not necessarily has to be a prediction task. We use it here as an example, but other applications (e.g., classification, regression etc.) are of course also possible.

Prediction with multiple models

- When you have a posterior $p(m_k|D)$ for the models, you don't *need* to choose one model for the prediction task. You can do prediction by **Bayesian model averaging**, which combines the predictive power from all models:

$$\begin{aligned} p(x|D) &= \sum_k \int p(x, \theta, m_k|D)\, \mathrm{d}\theta \\ &= \sum_k \int p(x|\theta, m_k)\, p(\theta|m_k, D)\, p(m_k|D)\, \mathrm{d}\theta \\ &= \sum_k \underbrace{p(m_k|D)}_{\substack{\text{model} \\ \text{posterior}}} \cdot \int \underbrace{p(\theta|m_k, D)}_{\substack{\text{parameter} \\ \text{posterior}}} \underbrace{p(x|\theta, m_k)}_{\substack{\text{data generating} \\ \text{distribution}}} \mathrm{d}\theta \end{aligned}$$

- Alternatively, if you do need to work with one model (e.g. due to computational resource constraints), you can for instance select the model with largest posterior $p(m_k|D)$ and use that model for prediction. This is called **Bayesian model selection**.

- Bayesian model averaging is the principal way to apply PT to machine learning. You don't throw away information by discarding lesser performant models, but rather use PT (marginalization of models) to compute

$$p(\text{what-I-am-interested-in} \,|\, \text{all available information})$$

  exactly.

We're Done!

- In principle, you now have the recipe in your hands now to solve all your prediction/classification/regression etc problems by the same method:

  1. specify a model
  2. train the model (by PT)
  3. evaluate the model (by PT); if not satisfied, goto 1
  4. apply the model (by PT)

- Crucially, there is no need to invent clever machine learning algorithms, and there is no need to invent a clever prediction algorithm nor a need to invent a model performance criterion. Instead, you propose a model and, from there on, you let PT reason about everything that you care about.

- Your problems are only of computational nature. Perhaps the integral to compute the evidence may not be analytically tractable, etc.

Bayesian Evidence as a Model Performance Criterion

- I'd like to convince you that Bayesian model evidence is an excellent criterion for assessing your model's performance. To do so, let us consider a decomposition that relates model evidence to other valued criteria such as **accuracy** and **model complexity**.

- Consider a model $p(x, \theta|m)$ and a data set $D = \{x_1, x_2, \ldots, x_N\}$.

- Given the data set D, the log-evidence for model m decomposes as follows (please check the derivation):

$$\begin{aligned} \underbrace{\log p(D|m)}_{\text{log-evidence}} &= \log \frac{p(D|\theta, m)p(\theta|m)}{p(\theta|D, m)} \qquad (\text{use Bayes rule}) \\ &= \underbrace{\log \frac{p(D|\theta, m)p(\theta|m)}{p(\theta|D, m)}}_{\substack{\text{evaluates to } \log p(D|m) \\ \text{therefore independent of } \theta}} \cdot \underbrace{\int p(\theta|D, m)\mathrm{d}\theta}_{\text{evaluates to } 1} \\ &= \int p(\theta|D, m) \cdot \underbrace{\log \frac{p(D|\theta, m)p(\theta|m)}{p(\theta|D, m)}}_{=\log p(D|m)} \mathrm{d}\theta \qquad (\text{move } \log p(D|m) \text{ into the integral}) \\ &= \underbrace{\int p(\theta|D, m) \log p(D|\theta, m) \mathrm{d}\theta}_{\text{accuracy (a.k.a. data fit)}} - \underbrace{\int p(\theta|D, m) \log \frac{p(\theta|D, m)}{p(\theta|m)} \mathrm{d}\theta}_{\text{complexity}} \end{aligned}$$

- The first term (data fit, also known as accuracy) measures how well the model predicts the data set D (as measured by $\log p(D|\theta, m)$), after having learned from the data (because we marginalize θ with the posterior $p(\theta|D, m)$). We want this term to be large (although only focussing on this term could lead to *overfitting*).

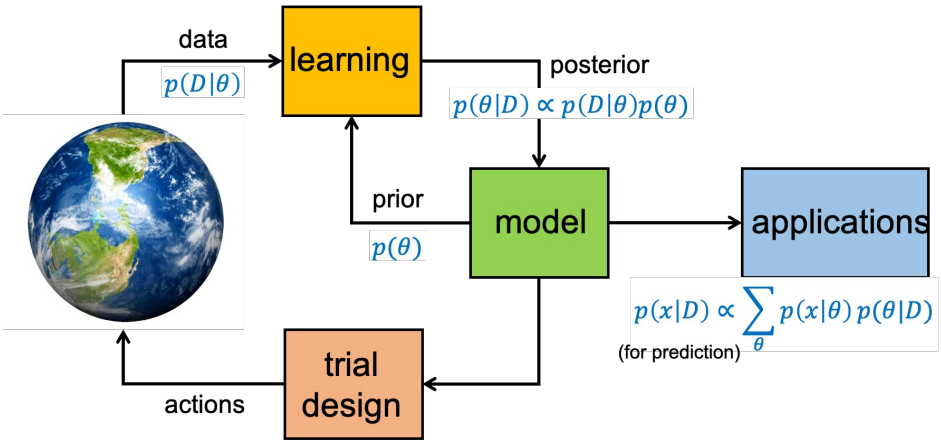
- The second term (complexity) quantifies the amount of information that the model absorbed through learning, i.e., by moving parameter beliefs from $p(\theta|m)$ to $p(\theta|D, m)$. Technically, this term is the Kullback-Leibler divergence between posterior and prior. We want this term to be small.

- The complexity term regularizes the Bayesian learning process automatically. If you prefer models with high Bayesian evidence, then you prefer models that get a good data fit without need to learn much from the data set. These types of models are said to *generalize* well, since they can be applied to different data sets without specific adaptations for each data set.

- ⇒ Bayesian learning automatically leads to models that generalize well. **There is no need for early stopping or validation data sets**. Just learn on the full data set and all behaves well.

Bayesian Machine Learning and the Scientific Method Revisited

- The Bayesian design process provides a unified framework for the Scientific Inquiry method. We can now add equations to the design loop. (Trial design to be discussed in Intelligent Agent lesson.)



Now Solve the Example Problem: Predicting a Coin Toss

- We observe a the following sequence of heads (h) and tails (t) when tossing the same coin repeatedly

$$D = \{hthhtth\}\,.$$

- What is the probability that heads comes up next? We solve this in the next slides ...

Coin toss example (1): Model Specification

- We observe a sequence of N coin tosses $D = \{x_1, \ldots, x_N\}$ with n heads.
- Let us denote outcomes by

$$x_k = \begin{cases} h & \text{if heads comes up} \\ t & \text{if tails} \end{cases}$$

Likelihood

- Assume a **Bernoulli** distributed variable $p(x_k = h|\mu) = \mu$ for a single coin toss, leading to

$$p(x_k|\mu) = \mu^x k(1-\mu)^{1-x_k}\,.$$

- Assume n times heads were thrown out of a total of N throws. The likelihood function then follows a a **binomial** distribution :

$$p(D|\mu) = \prod_{k=1}^{N} p(x_k|\mu) = \mu^n (1-\mu)^{N-n}$$

Prior

- Assume the prior beliefs for μ are governed by a **beta distribution**

$$p(\mu) = \text{Beta}(\mu|\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\mu^{\alpha-1}(1-\mu)^{\beta-1}$$

where the Gamma function is sort of a generalized factorial function. In particular, if α, β are integers, then

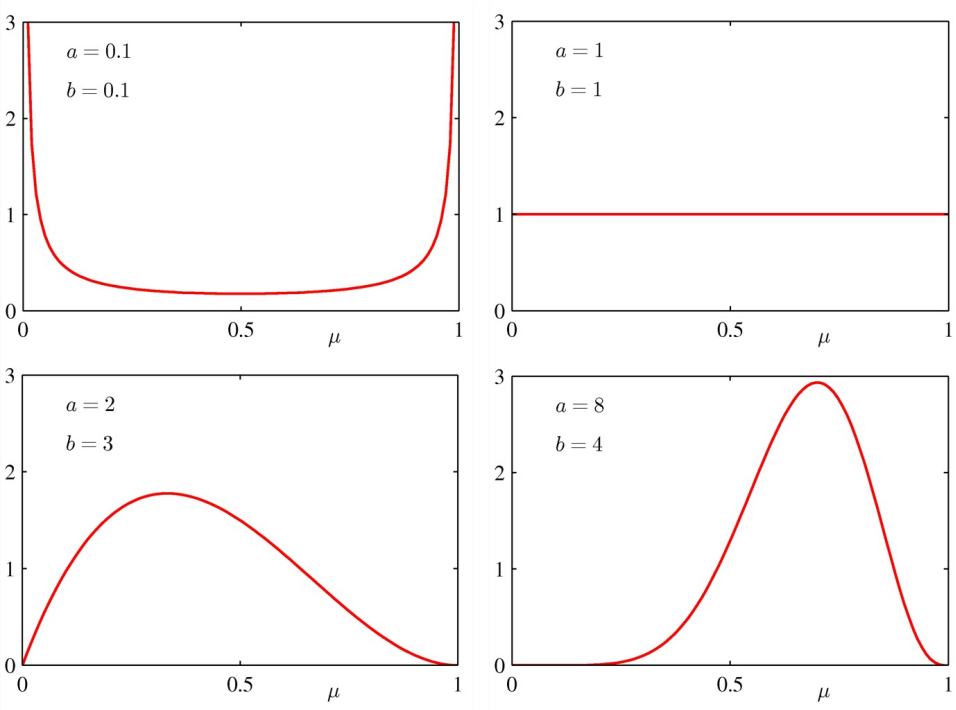
$$\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)(\Gamma(\beta)} = \frac{(\alpha+\beta-1)!}{(\alpha-1)!\,(\beta-1)!}$$

- A *what* distribution? Yes, the **beta distribution** is a **conjugate prior** for the binomial distribution, which means that

$$\underbrace{\text{beta}}_{\text{posterior}} \propto \underbrace{\text{binomial}}_{\text{likelihood}} \times \underbrace{\text{beta}}_{\text{prior}}$$

so we get a closed-form posterior.

- α and β are called **hyperparameters**, since they parameterize the distribution for another parameter (μ). E.g., $\alpha = \beta = 1$ (uniform).



- (Bishop Fig.2.2). Plots of the beta distribution $\text{Beta}(\mu|a,b)$ as a function of μ for various values of the hyperparameters a and b.

Coin toss example (2): Parameter estimation

- Infer posterior PDF over μ (and evidence) through Bayes rule

$$\begin{aligned} p(\mu|D) \cdot p(D) &= p(D|\mu) \cdot p(\mu) \\ &= \left[\mu^n (1-\mu)^{N-n}\right] \cdot \left[\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \mu^{\alpha-1}(1-\mu)^{\beta-1}\right] \\ &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \mu^{n+\alpha-1}(1-\mu)^{N-n+\beta-1} \\ &= \underbrace{\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(N+\alpha)\Gamma(N-n+\beta)}{\Gamma(N+\alpha+\beta)}}_{\text{evidence } p(D)} \cdot \left[\underbrace{\frac{\Gamma(N+\alpha+\beta)}{\Gamma(n+\alpha)\Gamma(N-n+\beta)} \mu^{n+\alpha-1}(1-\mu)^{N-n+\beta-1}}_{\text{posterior } p(\mu|D)=\text{Beta}(\mu|n+\alpha,N-n+\beta)}\right] \end{aligned}$$

hence the posterior is also beta-distributed as

$$p(\mu|D) = \text{Beta}(\mu|\, n+\alpha, N-n+\beta)$$

Coin toss example (3): Model Evaluation

- It follow from the above calculation that the evidence for model m can be analytically expressed as

$$p(D|m) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(N+\alpha)\Gamma(N-n+\beta)}{\Gamma(N+\alpha+\beta)}$$

- The model evidence is a scalar. The absolute value is not important. However, you may want to compare the model evidence of this model to the evidence for another model on the same data set.

Coin Toss Example (4): Prediction

- Once we have accepted a model, let's apply it to the application, in this case, predicting future observations.

- Marginalize over the parameter posterior to get the predictive PDF for a new coin toss $x_\bullet$, given the data D,

$$
\begin{aligned}
p(x_\bullet = h|D) &= \int_0^1 p(x_\bullet = h|\mu)\, p(\mu|D)\, \mathrm{d}\mu \\
&= \int_0^1 \mu \times \mathrm{Beta}(\mu|\, n+\alpha, N-n+\beta)\, \mathrm{d}\mu \\
&= \frac{n+\alpha}{N+\alpha+\beta}
\end{aligned}
$$

- This result is known as **Laplace's rule of succession**.
- The above integral computes the mean of a beta distribution, which is given by $\mathbb{E}[x] = \frac{a}{a+b}$ for $x \sim \mathrm{Beta}(a, b)$, see wikipedia.
- Finally, we're ready to solve our example problem: for $D = \{hthhtth\}$ and uniform prior ($\alpha = \beta = 1$), we get

$$
p(x_\bullet = h|D) = \frac{n+1}{N+2} = \frac{4+1}{7+2} = \frac{5}{9}
$$

Coin Toss Example: What did we learn?

- What did we learn from the data? Before seeing any data, we think that

$$
p(x_\bullet = h) = p(x_\bullet = h|D)|_{n=N=0} = \frac{\alpha}{\alpha+\beta}.
$$

- Hence, α and β are prior pseudo-counts for heads and tails respectively.
- After the N coin tosses, we think that $p(x_\bullet = h|D) = \frac{n+\alpha}{N+\alpha+\beta}$.
- Note the following decomposition

$$
\begin{aligned}
p(x_\bullet = h|\, D) &= \frac{n+\alpha}{N+\alpha+\beta} = \frac{n}{N+\alpha+\beta} + \frac{\alpha}{N+\alpha+\beta} \\
&= \frac{N}{N+\alpha+\beta} \cdot \frac{n}{N} + \frac{\alpha+\beta}{N+\alpha+\beta} \cdot \frac{\alpha}{\alpha+\beta} \\
&= \underbrace{\underbrace{\frac{\alpha}{\alpha+\beta}}_{\substack{\text{prior} \\ \text{prediction}}} + \underbrace{\frac{N}{N+\alpha+\beta}}_{\text{gain}} \cdot \Big(\underbrace{\underbrace{\frac{n}{N}}_{\substack{\text{data-based} \\ \text{prediction}}} - \underbrace{\frac{\alpha}{\alpha+\beta}}_{\substack{\text{prior} \\ \text{prediction}}}}_{\text{prediction error}} \Big)}_{\text{correction}}
\end{aligned}
$$

- Note that, since $0 \leq \underbrace{\frac{N}{N+\alpha+\beta}}_{\text{gain}} < 1$, the Bayesian prediction lies between (fuses) the prior and data-based predictions. The data plays the role of "correcting" the prior prediction.
- For large N, the gain goes to 1 and $p(x_\bullet = h|D)|_{N\to\infty} \to \frac{n}{N}$ goes to the data-based prediction (the observed relative frequency).

Code Example: Bayesian evolution for the coin toss

- Next, we code an example for a sequence of coin tosses, where we assume that the true coin generates data $x_n \in \{0, 1\}$ by a Bernoulli distribution:

$$
p(x_n|\mu = 0.4) = 0.4^{x_n} \cdot 0.6^{1-x_n}
$$

- So, this coin is biased!
- In order predict the outcomes of future coin tosses, we'll use two models m_1 and m_2.
- Both models have the same data generating distribution (also Bernoulli)

$$
p(x_n|\mu, m_1) = p(x_n|\mu, m_2) = \mu^{x_n}(1-\mu)^{1-x_n}
$$

but they have different priors:

$$
\begin{aligned}
p(\mu|m_1) &= \mathrm{Beta}(\mu|\alpha = 1, \beta = 1) \\
p(\mu|m_2) &= \mathrm{Beta}(\mu|\alpha = 15, \beta = 1)
\end{aligned}
$$

- Which model is better?
- For both models, we will report as a function of the total number of coin tosses, the posteriors

$$
p(\mu|D, m_1) \text{ and } p(\mu|D, m_2),
$$

and the Bayes factor in decibels

$$
B_{12} = log_{10}\left(\frac{p(D|m_1)}{p(D|m_2)}\right).
$$

In [1]:

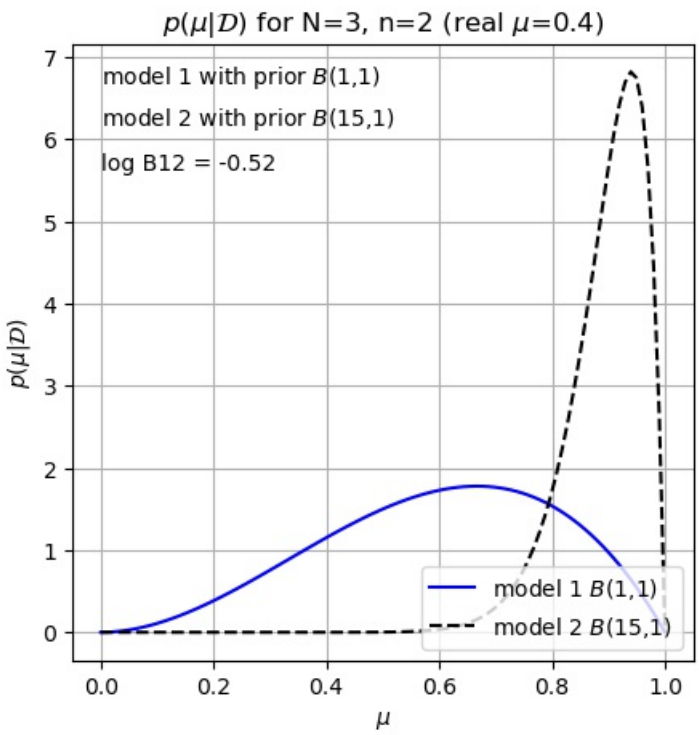
```
# computes log10 of Gamma function
function log10gamma(int)
    if int == 1 || int == 2
        return 1
    end
    return sum(log10(ii) for ii in 2:int-1)
end

# compute posterior and log-evidence
function execute_bayes(N,n,α,β)
# N, n is total # tosses and heads counts
# \alpha, \beta are parameters for Beta prior
    posterior = Beta( α + n, β + (N-n) )
    logevidence = log10gamma(α+β) - log10gamma(α) - log10gamma(β) + log10gamma(N+α) + log10gamma(N-n+β) - log10gamma(N+α+
    return posterior, logevidence
end;
```

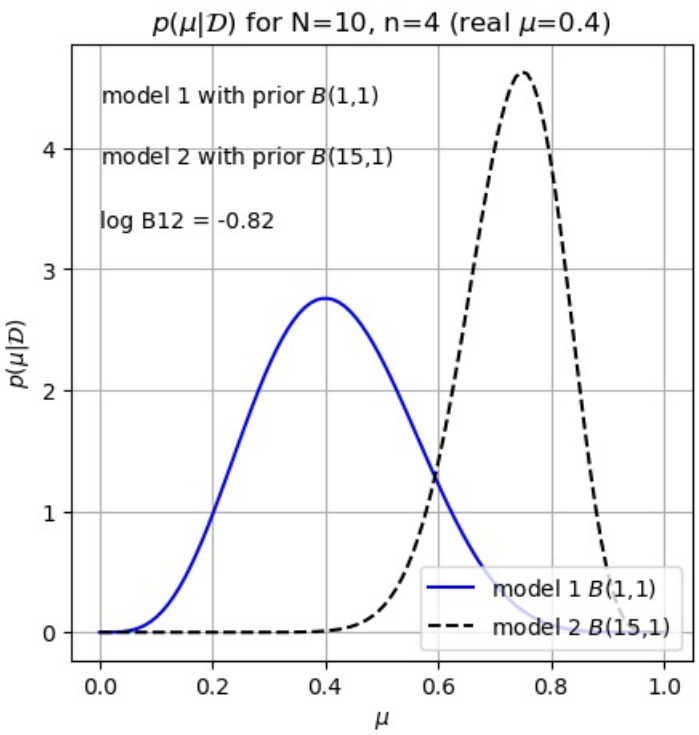
In [2]:

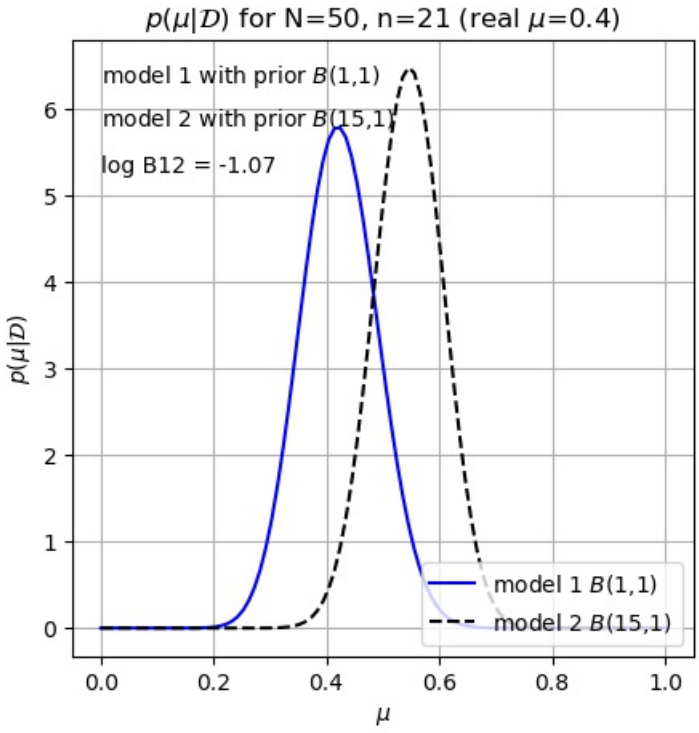
```
using Distributions

    # specify model parameters
    μ  = 0.4;
    α_m1 = 1; β_m1 = 1;
    α_m2 = 15; β_m2 = 1;

    # do experiment and update model
    max_ntosses = 192
    samples = rand(max_ntosses) .<= μ # Flip 192 coins

    posterior_m1 = Array{Distribution}(undef,max_ntosses)
    posterior_m2 = Array{Distribution}(undef,max_ntosses)
    logevidence_m1 = Array{Float64}(undef,max_ntosses)
    logevidence_m2 = Array{Float64}(undef,max_ntosses)
    logBF12 = Array{Float64}(undef,max_ntosses)

    for ntosses = 1:1:max_ntosses
        nheads = sum(samples[1:ntosses]) # Count number of heads in first N flips
        posterior_m1[ntosses], logevidence_m1[ntosses] = execute_bayes( ntosses, nheads, α_m1, β_m1 )
        posterior_m2[ntosses], logevidence_m2[ntosses] = execute_bayes( ntosses, nheads, α_m2, β_m2 )
        logBF12[ntosses] = logevidence_m1[ntosses] - logevidence_m2[ntosses]
    end
```

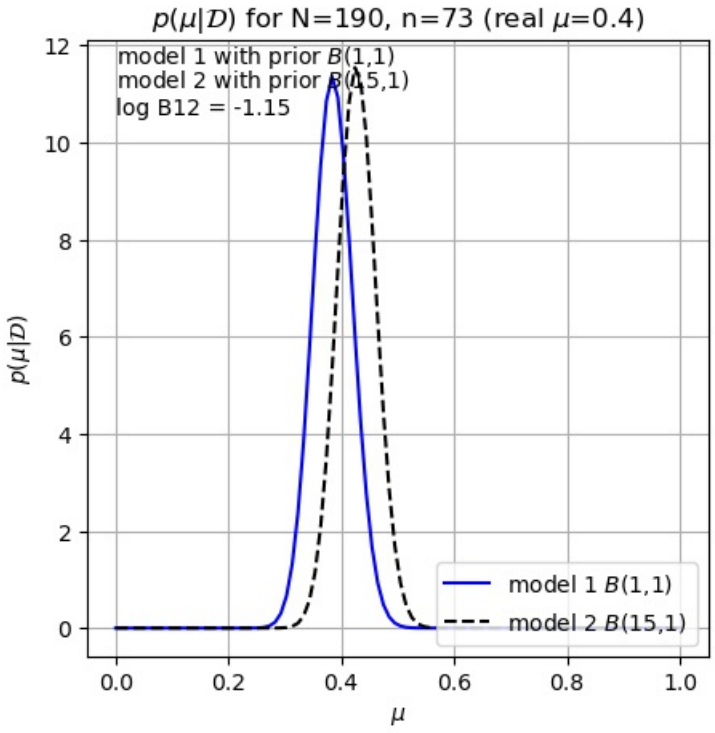
In [3]:

```
using PyPlot

    fig = figure("Posterior distributions", figsize=(11,11));
    range_grid = range(0.0, stop=1.0, length=100)
    ax1 = fig.add_subplot(2,2,1);
    ax2 = fig.add_subplot(2,2,2);
    ax3 = fig.add_subplot(2,2,3);
    ax4 = fig.add_subplot(2,2,4);

    function plot_distribution(position, ntosses, posterior_m1, posterior_m2, logBF12)
        plt.subplot(position);
        plot(range_grid,pdf.(posterior_m1[ntosses],range_grid), "b-");
        plt.subplot(position);
        plot(range_grid,pdf.(posterior_m2[ntosses],range_grid), "k--");

        xlabel(L"\mu"); ylabel(L"p(\mu|\mathcal{D})"); grid()
        title(L"p(\mu|\mathcal{D})"*" for N=$(ntosses), n=$(sum(samples[1:ntosses])) (real \$\\mu\$=$(μ))")
        legend(["model 1 "*L"B("*string(α_m1)*","*string(β_m1)*")","model 2 "*L"B("*string(α_m2)*","*string(β_m2)*")"], l
        ymax = max(maximum((pdf.(posterior_m1[ntosses],range_grid))),maximum((pdf.(posterior_m2[ntosses],range_grid))))
        text(0, ymax-.5,"model 1 with prior "*L"B("*string(α_m1)*","*string(β_m1)*")\n")
        text(0, ymax-1, "model 2 with prior "*L"B("*string(α_m2)*","*string(β_m2)*")\n")
        text(0, ymax-1.5, "log B12 = $(round(logBF12[ntosses], digits=2))\n")
    end

    function plot_all_distributions(posterior_m1,posterior_m2,logBF12)
        ntosses = [3,10,50,190];
        plot_distribution(221, ntosses[1], posterior_m1, posterior_m2, logBF12)
        plot_distribution(222, ntosses[2], posterior_m1, posterior_m2, logBF12)
        plot_distribution(223, ntosses[3], posterior_m1, posterior_m2, logBF12)
        plot_distribution(224, ntosses[4], posterior_m1, posterior_m2, logBF12)
    end

    plot_all_distributions( posterior_m1, posterior_m2, logBF12 );
```









⇒ With more data, the relevance of the prior diminishes!

From Posterior to Point-Estimate

- In the example above, Bayesian parameter estimation and prediction were tractable in closed-form. This is often not the case. We will need to approximate some of the computations.
- Recall Bayesian prediction

$$p(x|D) = \int p(x|\theta)p(\theta|D)\,\mathrm{d}\theta$$

- If we approximate posterior $p(\theta|D)$ by a delta function for one 'best' value $\hat{\theta}$, then the predictive distribution collapses to

$$p(x|D) = \int p(x|\theta)\,\delta(\theta-\hat{\theta})\,\mathrm{d}\theta = p(x|\hat{\theta})$$

- This is just the data generating distribution $p(x|\theta)$ evaluated at $\theta = \theta$, which is easy to evaluate.
- The next question is how to get the parameter estimate θ? (See next slide).

Some Well-known Point-Estimates

- **Bayes estimate** (the mean of the posterior)

$$\theta_{bayes} = \int \theta p\left(\theta|D\right) \mathrm{d}\theta$$

- **Maximum A Posteriori** (MAP) estimate

$$\theta_{\text{map}} = \arg\max_{\theta} p\left(\theta|D\right) = \arg\max_{\theta} p\left(D|\theta\right)\, p\left(\theta\right)$$

- **Maximum Likelihood** (ML) estimate

$$\theta_{ml} = \arg\max_{\theta} p\left(D|\theta\right)$$

  - Note that Maximum Likelihood is MAP with uniform prior
  - ML is the most common approximation to the full Bayesian posterior.

Bayesian vs Maximum Likelihood Learning

Consider the task: predict a datum x from an observed data set D.

	Bayesian	**Maximum Likelihood**
1. **Model Specification**	Choose a model m with data generating distribution $p(x|\theta, m)$ and parameter prior $p(\theta|m)$	Choose a model m with same data generating distribution $p(x|\theta, m)$. No need for priors.
2. **Learning**	use Bayes rule to find the parameter posterior, $p(\theta|D) \propto p(D|\theta)p(\theta)$	By Maximum Likelihood (ML) optimization, $\theta = \arg\max_{\theta} p(D|\theta)$
3. **Prediction**	$p(x|D) = \int p(x|\theta)p(\theta|D)\,\mathrm{d}\theta$	$p(x|D) = p(x|\theta)$

Report Card on Maximum Likelihood Estimation

- Maximum Likelihood (ML) is MAP with uniform prior. MAP is sometimes called a 'penalized' ML procedure:

$$\hat{\theta}_{map} = \arg\max_{\theta}\{\underbrace{\log p\left(D|\theta\right)}_{\text{log-likelihood}} + \underbrace{\log p\left(\theta\right)}_{\text{penalty}}\}$$

- (good!). ML works rather well if we have a lot of data because the influence of the prior diminishes with more data.
- (good!). Computationally often do-able. Useful fact that makes the optimization easier (since $\log$ is monotonously increasing):

$$\arg\max_{\theta} \log p(D|\theta) = \arg\max_{\theta} p(D|\theta)$$

- (bad). Cannot be used for model comparison! When doing ML estimation, the Bayesian model evidence always evalutes to zero because the prior probability mass under the likelihood function goes to zero. Therefore, when doing ML estimation, Bayesian model evidence cannot be used to evaluate model performance:

$$\begin{aligned}
\underbrace{p(D|m)}_{\substack{\text{Bayesian}\\ \text{evidence}}} &= \int p(D|\theta) \cdot p(\theta|m)\,\mathrm{d}\theta \\
&= \lim_{(b-a)\to\infty} \int p(D|\theta) \cdot \text{Uniform}(\theta|a, b)\,\mathrm{d}\theta \\
&= \lim_{(b-a)\to\infty} \frac{1}{b-a} \underbrace{\int_a^b p(D|\theta)\,\mathrm{d}\theta}_{<\infty} \\
&= 0
\end{aligned}$$

  - In fact, this is a serious problem because evidence is fundamentally the correct criterion that follows from straighforward PT. In practice, when estimating parameters by maximum likelihood, we evaluate model performance by an ad hoc performance measure such as mean-squared-error on a testing data set.

Factor Graphs

Preliminaries

- Goal
  - Introduction to Forney-style factor graphs and message passing-based inference
- Materials
  - Mandatory
    - These lecture notes
    - Loeliger (2007), The factor graph approach to model based signal processing, pp. 1295-1302 (until section V)
  - Optional
    - Frederico Wadehn (2015), Probabilistic graphical models: Factor graphs and more video lecture (**highly recommended**)
  - References
    - Forney (2001), Codes on graphs: normal realizations

Why Factor Graphs?

- A probabilistic inference task gets its computational load mainly through the need for marginalization (i.e., computing integrals). E.g., for a model $p(x_1, x_2, x_3, x_4, x_5)$, the inference task $p(x_2|x_3)$ is given by

$$p(x_2|x_3) = \frac{p(x_2,x_3)}{p(x_3)} = \frac{\int\cdots\int p(x_1,x_2,x_3,x_4,x_5)\,\mathrm{d}x_1\mathrm{d}x_4\mathrm{d}x_5}{\int\cdots\int p(x_1,x_2,x_3,x_4,x_5)\,\mathrm{d}x_1\mathrm{d}x_2\mathrm{d}x_4\mathrm{d}x_5}$$

- Since these computations (integrals or sums) suffer from the "curse of dimensionality", we often need to solve a simpler problem in order to get an answer.

- Factor graphs provide a computationally efficient approach to solving inference problems **if the probabilistic model can be factorized**.

- Factorization helps. For instance, if $p(x_1, x_2, x_3, x_4, x_5) = p(x_1)p(x_2,x_3)p(x_4)p(x_5|x_4)$, then

$$p(x_2|x_3) = \frac{\int\cdots\int p(x_1)p(x_2,x_3)p(x_4)p(x_5|x_4)\,\mathrm{d}x_1\mathrm{d}x_4\mathrm{d}x_5}{\int\cdots\int p(x_1)p(x_2,x_3)p(x_4)p(x_5|x_4)\,\mathrm{d}x_1\mathrm{d}x_2\mathrm{d}x_4\mathrm{d}x_5} = \frac{p(x_2,x_3)}{\int p(x_2,x_3)\mathrm{d}x_2}$$

which is computationally much cheaper than the general case above.

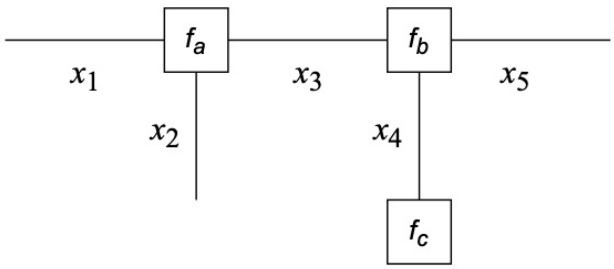
- In this lesson, we discuss how computationally efficient inference in *factorized* probability distributions can be automated by message passing-based inference in factor graphs.

Factor Graph Construction Rules

- Consider a function

$$f(x_1,x_2,x_3,x_4,x_5) = f_a(x_1,x_2,x_3)\cdot f_b(x_3,x_4,x_5)\cdot f_c(x_4)$$

- The factorization of this function can be graphically represented by a **Forney-style Factor Graph** (FFG):



- An FFG is an **undirected** graph subject to the following construction rules (Forney, 2001)

  1. A **node** for every factor;
  2. An **edge** (or **half-edge**) for every variable;
  3. Node $f_\bullet$ is connected to edge x **iff** variable x appears in factor $f_\bullet$.

- A **configuration** is an assigment of values to all variables.

- A configuration $\omega = (x_1, x_2, x_3, x_4, x_5)$ is said to be **valid** iff $f(\omega) \neq 0$

Equality Nodes for Branching Points

- Note that a variable can appear in maximally two factors in an FFG (since an edge has only two end points).
- Consider the factorization (where x_2 appears in three factors)

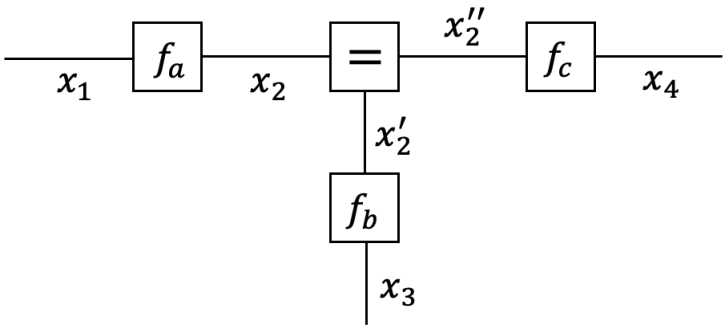
$$f(x_1, x_2, x_3, x_4) = f_a(x_1, x_2) \cdot f_b(x_2, x_3) \cdot f_c(x_2, x_4)$$

- For the factor graph representation, we will instead consider the function g, defined as

$$g(x_1, x_2, x_2', x_2'', x_3, x_4) = f_a(x_1, x_2) \cdot f_b(x_2', x_3) \cdot f_c(x_2'', x_4) \cdot f_=(x_2, x_2', x_2'')$$

where

$$f_=(x_2, x_2', x_2'') \triangleq \delta(x_2 - x_2')\,\delta(x_2 - x_2'')$$



- Note that through introduction of auxiliary variables X_2' and X_2'' and a factor $f_=(x_2, x_2', x_2'')$, each variable in g appears in maximally two factors.
- The constraint $f_=(x, x', x'')$ enforces that $X = X' = X''$ **for every valid configuration**.
- Since f is a marginal of g, i.e.,

$$f(x_1, x_2, x_3, x_4) = \iint g(x_1, x_2, x_2', x_2'', x_3, x_4)\, \mathrm{d}x_2' \mathrm{d}x_2''$$

it follows that any inference problem on f can be executed by a corresponding inference problem on g, e.g.,

$$\begin{aligned} f(x_1 \mid x_2) &\triangleq \frac{\iint f(x_1, x_2, x_3, x_4)\, \mathrm{d}x_3 \mathrm{d}x_4}{\int \cdots \int f(x_1, x_2, x_3, x_4)\, \mathrm{d}x_1 \mathrm{d}x_3 \mathrm{d}x_4} \\ &= \frac{\int \cdots \int g(x_1, x_2, x_2', x_2'', x_3, x_4)\, \mathrm{d}x_2' \mathrm{d}x_2'' \mathrm{d}x_3 \mathrm{d}x_4}{\int \cdots \int g(x_1, x_2, x_2', x_2'', x_3, x_4)\, \mathrm{d}x_1 \mathrm{d}x_2' \mathrm{d}x_2'' \mathrm{d}x_3 \mathrm{d}x_4} \\ &= g(x_1 \mid x_2) \end{aligned}$$

- $\Rightarrow$ **Any factorization of a global function f can be represented by a Forney-style Factor Graph**.

Probabilistic Models as Factor Graphs

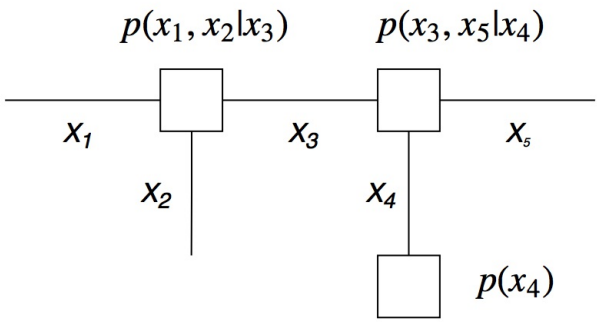
- FFGs can be used to express conditional independence (factorization) in probabilistic models.
- For example, the (previously shown) graph for $f_a(x_1, x_2, x_3) \cdot f_b(x_3, x_4, x_5) \cdot f_c(x_4)$ could represent the probabilistic model

$$p(x_1, x_2, x_3, x_4, x_5) = p(x_1, x_2 | x_3) \cdot p(x_3, x_5 | x_4) \cdot p(x_4)$$

where we identify

$$\begin{aligned} f_a(x_1, x_2, x_3) &= p(x_1, x_2 | x_3) \\ f_b(x_3, x_4, x_5) &= p(x_3, x_5 | x_4) \\ f_c(x_4) &= p(x_4) \end{aligned}$$

- This is the graph



Inference by Closing Boxes

- Factorizations provide opportunities to cut on the amount of needed computations when doing inference. In what follows, we will use FFGs to process these opportunities in an automatic way by message passing between the nodes of the graph.

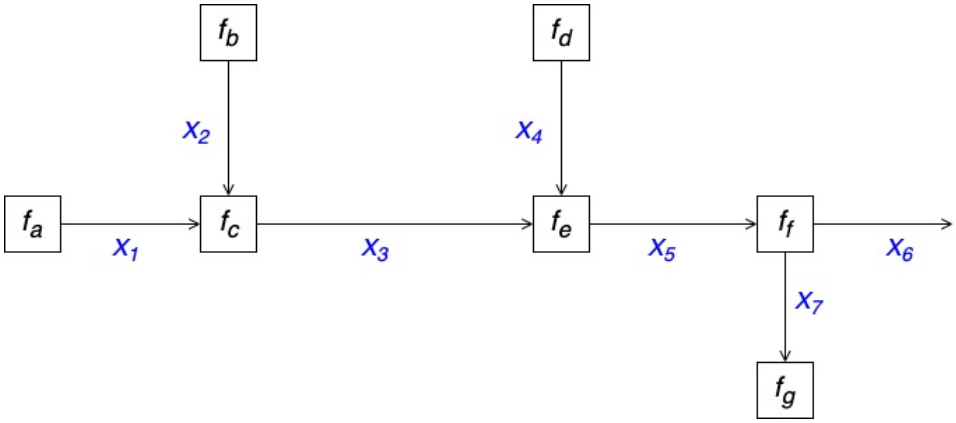
- Assume we wish to compute the marginal

$$f(x_3) \triangleq \sum_{x_1,x_2,x_4,x_5,x_6,x_7} f(x_1, x_2, \ldots, x_7)$$

  for a model f with given factorization

$$f(x_1, x_2, \ldots, x_7) = f_a(x_1) f_b(x_2) f_c(x_1, x_2, x_3) f_d(x_4) f_e(x_3, x_4, x_5) f_f(x_5, x_6, x_7) f_g(x_7)$$

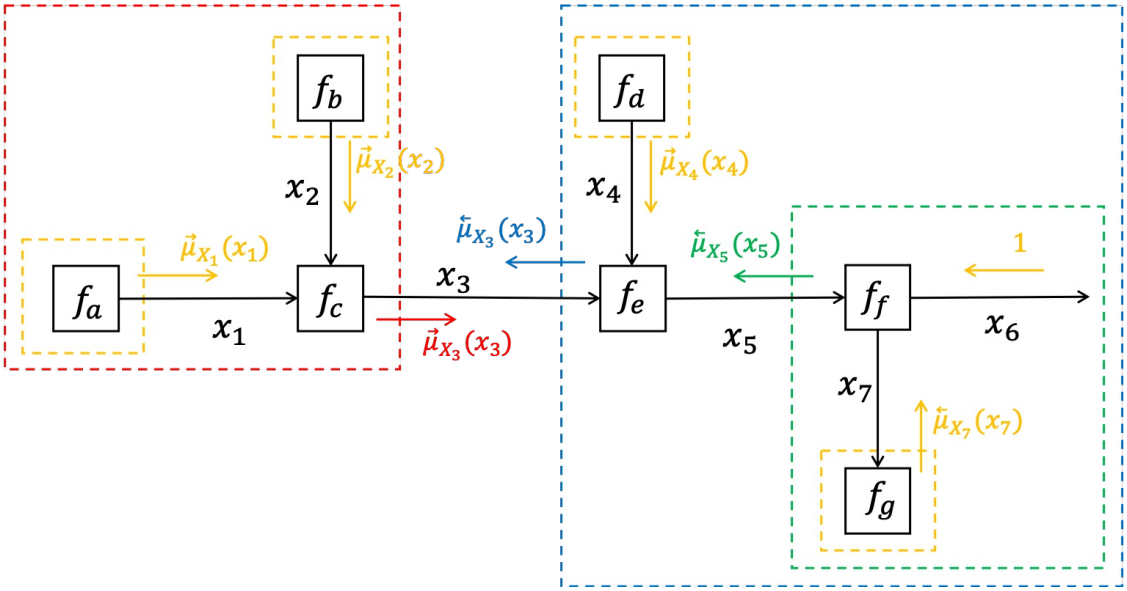
- The FFG for f is (we will discuss the usage of directed edges below):



- Due to the factorization and the Generalized Distributive Law, we can decompose this sum-of-products to the following product-of-sums:

$$f(x_3) = \underbrace{\left(\sum_{x_1,x_2} \underbrace{f_a(x_1)}_{\vec{\mu}_{X_1}(x_1)} \underbrace{f_b(x_2)}_{\vec{\mu}_{X_2}(x_2)} f_c(x_1, x_2, x_3) \right)}_{\vec{\mu}_{X_3}(x_3)} \cdot \underbrace{\left(\sum_{x_4,x_5} \underbrace{f_d(x_4)}_{\vec{\mu}_{X_4}(x_4)} f_e(x_3, x_4, x_5) \cdot \underbrace{\left(\sum_{x_6,x_7} f_f(x_5, x_6, x_7) \underbrace{f_g(x_7)}_{\overleftarrow{\mu}_{X_7}(x_7)} \right)}_{\overleftarrow{\mu}_{X_5}(x_5)} \right)}_{\overleftarrow{\mu}_{X_3}(x_3)}$$

  which is computationally (much) lighter than executing the full sum $\sum_{x_1,\ldots,x_7} f(x_1, x_2, \ldots, x_7)$

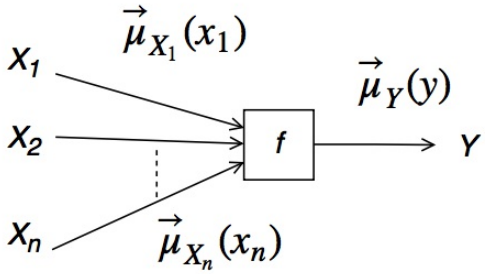


- Note that the new factor $\vec{\mu}_{X_3}(x_3)$ is obtained by multiplying all enclosed factors (f_a, f_b, f_c) by the red dashed box, followed by marginalization (summing) over all enclosed variables (x_1, x_2).
- This is the **Closing the Box**-rule, which is a general recipe for marginalization of latent variables (inside the box) and leads to a new factor that has the variables (edges) that cross the box as arguments. For instance, the argument of the remaining factor $\vec{\mu}_{X_3}(x_3)$ is the variable on the edge that crosses the red box (x_3).
- Hence, $\vec{\mu}_{X_3}(x_3)$ can be interpreted as a **message from the red box toward variable** x_3.
- We drew *directed edges* in the FFG in order to distinguish forward messages $\vec{\mu}_\bullet(\cdot)$ (in the same direction as the arrow of the edge) from backward messages $\overleftarrow{\mu}_\bullet(\cdot)$ (in opposite direction). This is just a notational convenience since an FFG is computationally an undirected graph.

Sum-Product Algorithm

- Closing-the-box can also be interpreted as a **message update rule** for an outgoing message from a node. For a node $f(y, x_1, \ldots, x_n)$ with incoming messages $\vec{\mu}_{X_1}(x_1), \vec{\mu}_{X_1}(x_1), \ldots, \vec{\mu}_{X_n}(x_n)$, the outgoing message is given by (Loeliger (2007), pg.1299):

$$\underbrace{\vec{\mu}_Y(y)}_{\substack{\text{outgoing}\\ \text{message}}} = \sum_{x_1,\ldots,x_n} \underbrace{\vec{\mu}_{X_1}(x_1)\cdots\vec{\mu}_{X_n}(x_n)}_{\substack{\text{incoming}\\ \text{messages}}} \cdot \underbrace{f(y, x_1, \ldots, x_n)}_{\substack{\text{node}\\ \text{function}}}$$



- This is called the **Sum-Product Message** (SPM) update rule. (Look at the formula to understand why it's called the SPM update rule).
- Just as a final note, inference by sum-product message passing is much like replacing the sum-of-products

$$ac + ad + bc + bd$$

by the following product-of-sums:

$$(a + b)(c + d)\,.$$

Which of these two computations is cheaper to execute?

- Note that all SPM update rules can be computed from information that is **locally available** at each node.
- If the factor graph for a function f has **no cycles** (i.e., the graph is a tree), then the marginal $f(x_3) = \sum_{x_1,x_2,x_4,x_5,x_6,x_7} f(x_1, x_2, \ldots, x_7)$ is given by multiplying the

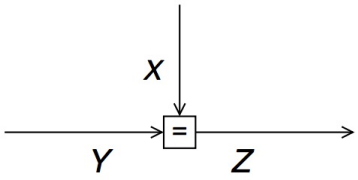
forward and backward messages on that edge:

$$\boxed{f(x_3) = \vec{\mu}_{X_3}(x_3) \cdot \overleftarrow{\mu}_{X_3}(x_3)}$$

- It follows that the marginal $f(x_3) = \sum_{x_1,x_2,x_4,x_5,x_6,x_7} f(x_1, x_2, \ldots, x_7)$ can be efficiently computed through sum-product messages. Executing inference through SP message passing is called the **Sum-Product Algorithm** (or alternatively, the **belief propagation** algorithm).

Sum-Product Messages for the Equality Node

- As an example, let´s evaluate the SP messages for the **equality node** $f_=(x, y, z) = \delta(z-x)\delta(z-y)$:



$$\begin{aligned}\vec{\mu}_Z(z) &= \iint \vec{\mu}_X(x)\,\vec{\mu}_Y(y)\,\delta(z-x)\delta(z-y)\,\mathrm{d}x\mathrm{d}y \\ &= \vec{\mu}_X(z) \int \vec{\mu}_Y(y)\,\delta(z-y)\,\mathrm{d}y \\ &= \vec{\mu}_X(z)\,\vec{\mu}_Y(z)\end{aligned}$$

- By symmetry, this also implies (for the same equality node) that

$$\begin{aligned}\overleftarrow{\mu}_X(x) &= \vec{\mu}_Y(x)\,\overleftarrow{\mu}_Z(x) \quad \text{and} \\ \overleftarrow{\mu}_Y(y) &= \vec{\mu}_X(y)\,\overleftarrow{\mu}_Z(y)\,.\end{aligned}$$

- Let us now consider the case of Gaussian messages $\vec{\mu}_X(x) = \mathcal{N}(x|\vec{m}_X, \vec{V}_X)$, $\vec{\mu}_Y(y) = \mathcal{N}(y|\vec{m}_Y, \vec{V}_Y)$ and $\vec{\mu}_Z(z) = \mathcal{N}(z|\vec{m}_Z, \vec{V}_Z)$. Let´s also define the precision matrices $\vec{W}_X \triangleq \vec{V}_X^{-1}$ and similarly for Y and Z. Then applying the SP update rule leads to multiplication of two Gaussian distributions, resulting in

$$\begin{aligned}\vec{W}_Z &= \vec{W}_X + \vec{W}_Y \\ \vec{W}_Z\vec{m}_z &= \vec{W}_X\vec{m}_X + \vec{W}_Y\vec{m}_Y\end{aligned}$$

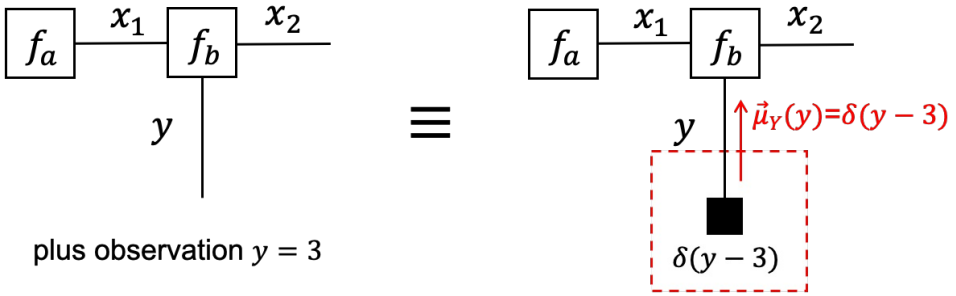
- It follows that **message passing through an equality node is similar to applying Bayes rule**, i.e., fusion of two information sources. Does this make sense?

Message Passing Schedules

- In a tree graph, start with messages from the terminals and keep passing messages through the internal nodes towards the "target" variable (x_3 in above problem) until you have both the forward and backward message for the target variable.
- In a tree graph, if you continue to pass messages throughout the graph, the Sum-Product Algorithm computes **exact** marginals for all hidden variables.
- If the graph contains cycles, we have in principle an infinite tree by "unrolling" the graph. In this case, the SP Algorithm is not guaranteed to find exact marginals. In practice, if we apply the SP algorithm for just a few iterations ("unrolls"), then we often find satisfying approximate marginals.

Terminal Nodes and Processing Observations

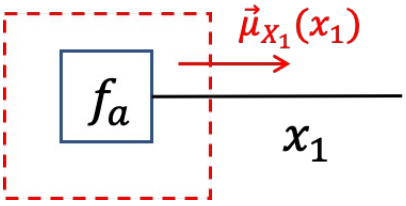
- We can use terminal nodes to represent observations, e.g., add a factor $f(y) = \delta(y-3)$ to terminate the half-edge for variable Y if $y=3$ is observed.



- Terminal nodes that carry observations are denoted by small black boxes.
- The message out of a **terminal node** (attached to only 1 edge) is the factor itself. For instance, closing a box around terminal node $f_a(x_1)$ would lead to

$$\vec{\mu}_{X_1}(x_1) \triangleq \sum_{\substack{\text{enclosed}\\ \text{variables}}} \prod_{\substack{\text{enclosed}\\ \text{factors}}} f_a(x_1) = f_a(x_1)$$

since there are no enclosed variables.



- The message from a half-edge is 1 (one). You can verify this by imagining that a half-edge x can be terminated by a node function $f(x) = 1$ without affecting any inference issue.

Automating Bayesian Inference by Message Passing

- The foregoing message update rules can be worked out in closed-form and put into tables (e.g., see Tables 1 through 6 in Loeliger (2007) for many standard factors such as essential probability distributions and operations such as additions, fixed-gain multiplications and branching (equality nodes).
- In the optional slides below, we have worked out a few of these update rules, eg, for the equality node, the addition node and the multiplication node.
- If the update rules for all node types in a graph have been tabulated, then inference by message passing comes down to executing a set of table-lookup operations, thus creating a completely **automatable Bayesian inference framework**.
- In our research lab BIASlab (FLUX 7.060), we are developing ReactiveMP.jl, which is a (Julia) toolbox for automating Bayesian inference by message passing in a factor graph.

Example: Bayesian Linear Regression by Message Passing

- Assume we want to estimate some function $f : \mathbb{R}^D \to \mathbb{R}$ from a given data set $D = \{(x_1, y_1), \ldots, (x_N, y_N)\}$, with model assumption $y_i = f(x_i) + \epsilon_i$.

model specification

- We will assume a linear model with white Gaussian noise and a Gaussian prior on the coefficients w:

$$\begin{aligned} y_i &= w^T x_i + \epsilon_i \\ \epsilon_i &\sim \mathcal{N}(0, \sigma^2) \\ w &\sim \mathcal{N}(0, \Sigma) \end{aligned}$$

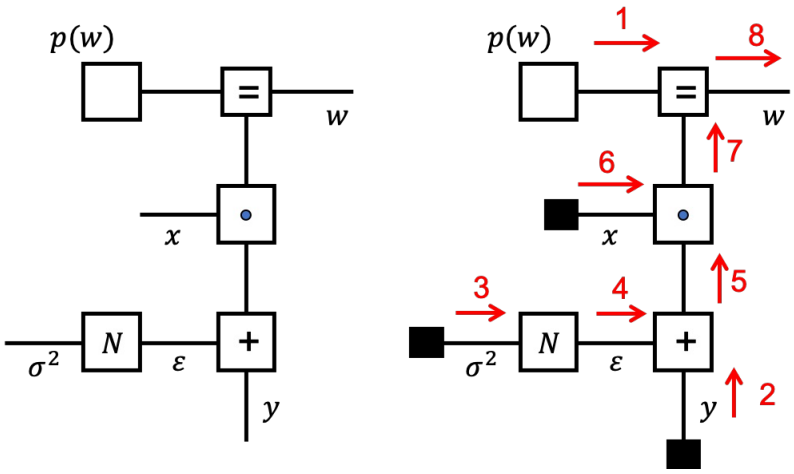
or equivalently

$$\begin{aligned} p(w, \epsilon, D) &= \overbrace{p(w)}^{\text{weight prior}} \prod_{i=1}^{N} \overbrace{p(y_i \mid x_i, w, \epsilon_i)}^{\text{regression model}} \overbrace{p(\epsilon_i)}^{\text{noise model}} \\ &= \mathcal{N}(w \mid 0, \Sigma) \prod_{i=1}^{N} \delta(y_i - w^T x_i - \epsilon_i) \mathcal{N}(\epsilon_i \mid 0, \sigma^2) \end{aligned}$$

Inference (parameter estimation)

- We are interested in inferring the posterior $p(w \mid D)$. We will execute inference by message passing on the FFG for the model.

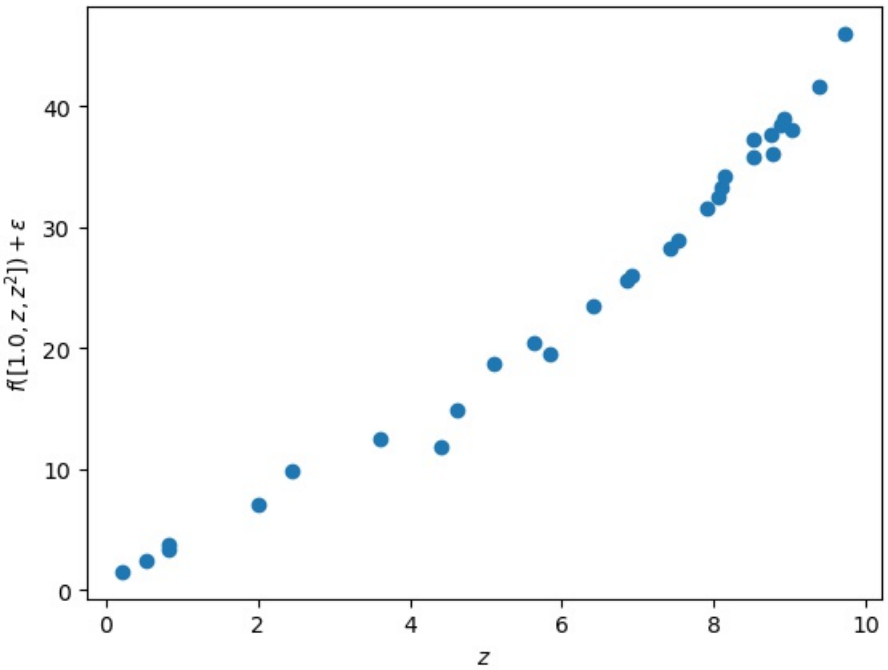
- The left figure shows the factor graph for this model.
- The right figure shows the message passing scheme.



CODE EXAMPLE

Let's solve this problem by message passing-based inference with Julia's FFG toolbox ReactiveMP.

In [1]:
```
using PyPlot, LinearAlgebra

# Parameters
Σ = 1e5 * Diagonal(I,3) # Covariance matrix of prior on w
σ2 = 2.0                # Noise variance

# Generate data set
w = [1.0; 2.0; 0.25]
N = 30
z = 10.0*rand(N)
x_train = [[1.0; z; z^2] for z in z] # Feature vector x = [1.0; z; z^2]
f(x) = (w'*x)[1]
y_train = map(f, x_train) + sqrt(σ2)*randn(N) # y[i] = w' * x[i] + ϵ
scatter(z, y_train); xlabel(L"z"); ylabel(L"f([1.0, z, z^2]) + \epsilon");
```

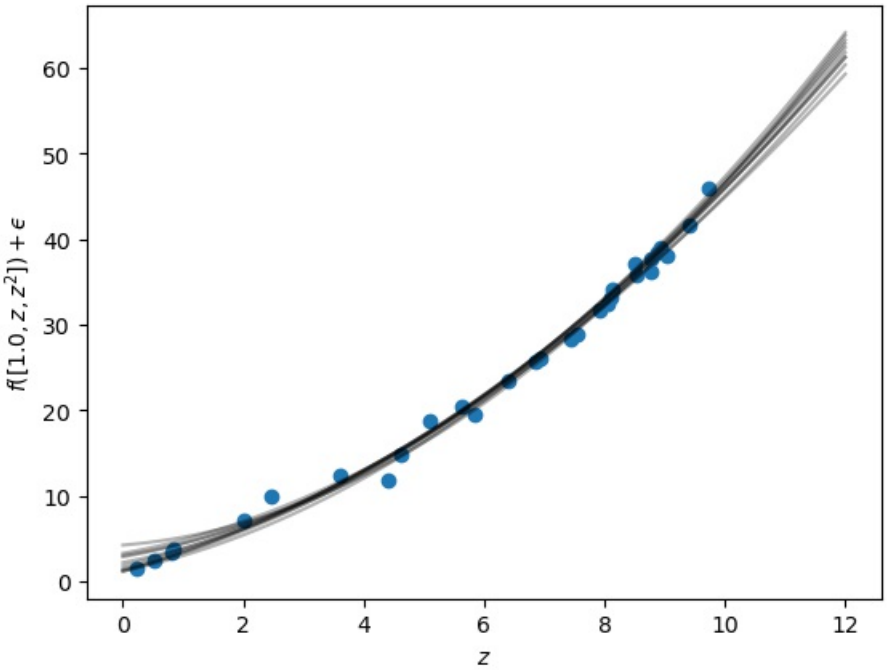


Now build the factor graph in ReactiveMP, perform sum-product message passing and plot results (mean of posterior).

In [2]:
```
# Build model
using ReactiveMP, GraphPPL, Random, Distributions
@model function linear_regression(N, Σ, σ2)

    w ~ MvNormalMeanCovariance(constvar(zeros(3)),constvar(Σ))

    x = datavar(Vector{Float64}, N)
    y = datavar(Float64, N)
```

```
        for i in 1:N
            y[i] ~ NormalMeanVariance(dot(w , x[i]), σ2)
        end
        return w, x, y
    end
    # Run message passing algorithm
    results = inference(
        model = Model(linear_regression, length(x_train), Σ, σ2 ),
        data  = (y = y_train, x = x_train),
        returnvars   = (w = KeepLast()),
        iterations   = 20
    );
    # Plot result
    w = results.posteriors[:w]
    println("Posterior distribution of w: $(w)")
    scatter(z, y_train); xlabel(L"z"); ylabel(L"f([1.0, z, z^2]) + \epsilon");
    z_test = collect(0:0.2:12)
    x_test = [[1.0; z; z^2] for z in z_test]
    for i=1:10
        w_sample = rand(results.posteriors[:w])
        f_est(x) = (w_sample'*x)[1]
        plot(z_test, map(f_est, x_test), "k-", alpha=0.3);
    end
```

```
Posterior distribution of w: MvNormalWeightedMeanPrecision(
xi: [369.9099796750249, 2836.80417217466, 23017.575289327015]
Λ: [15.000010000030002 92.01635691066768 693.7622183271991; 92.01635691066768 693.7622283272291 5551.6571006883; 693.7622
183271991 5551.6571006883 45815.28545656639]
)
```



Final thoughts: Modularity and Abstraction

- The great Michael Jordan (no, not this one, but this one), wrote:

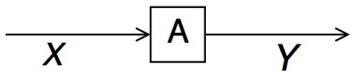
  > I basically know of two principles for treating complicated systems in simple ways: the first is the principle of **modularity** and the second is the principle of **abstraction**. I am an apologist for computational probability in machine learning because I believe that probability theory implements these two principles in deep and intriguing ways — namely through factorization and through averaging. Exploiting these two mechanisms as fully as possible seems to me to be the way forward in machine learning. — Michael Jordan, 1997 (quoted in Fre98).

- Factor graphs realize these ideas nicely, both visually and computationally.
- Visually, the modularity of conditional independencies in the model are displayed by the graph structure. Each node hides internal complexity and by closing-the-box, we can hierarchically move on to higher levels of abstraction.
- Computationally, message passing-based inference uses the Distributive Law to avoid any unnecessary computations.

OPTIONAL SLIDES

Sum-Product Messages for Multiplication Nodes

- Next, let us consider a **multiplication** by a fixed (invertible matrix) gain $f_A(x,y) = \delta(y - Ax)$



$$\vec{\mu}_Y(y) = \int \vec{\mu}_X(x)\, \delta(y - Ax)\, \mathrm{d}x = \vec{\mu}_X(A^{-1}y)\,.$$

- For a Gaussian message input message $\vec{\mu}_X(x) = \mathcal{N}(x|\vec{m}_X, \vec{V}_X)$, the output message is also Gaussian with

$$\vec{m}_Y = A\vec{m}_X \text{ , and } \vec{V}_Y = A\vec{V}_X A^T$$

since

$$\begin{aligned}
\vec{\mu}_Y(y) &= \vec{\mu}_X(A^{-1}y) \\
&\propto \exp\left(-\frac{1}{2}\left(A^{-1}y - \vec{m}_X\right)^T \vec{V}_X^{-1}\left(A^{-1}y - \vec{m}_X\right)\right) \\
&= \exp\big(-\frac{1}{2}\left(y - A\vec{m}_X\right)^T \underbrace{A^{-T}\vec{V}_X^{-1}A^{-1}}_{(A\vec{V}_X A^T)^{-1}}\left(y - A\vec{m}_X\right)\big) \\
&\propto \mathcal{N}(y|A\vec{m}_X, A\vec{V}_X A^T)\,.
\end{aligned}$$

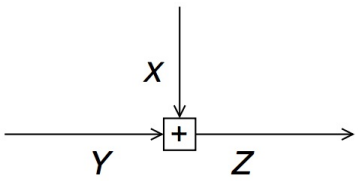
- **Excercise**: Proof that, for the same factor $\delta(y - Ax)$ and Gaussian messages, the (backward) sum-product message $\overleftarrow{\mu}_X$ is given by

$$\begin{aligned}
\overleftarrow{\xi}_X &= A^T\overleftarrow{\xi}_Y \\
\overleftarrow{W}_X &= A^T\overleftarrow{W}_Y A
\end{aligned}$$

where $\overleftarrow{\xi}_X \triangleq \overleftarrow{W}_X\overleftarrow{m}_X$ and $\overleftarrow{W}_X \triangleq \overleftarrow{V}_X^{-1}$ (and similarly for Y).

Code example: Gaussian forward and backward messages for the Addition node

Let's calculate the Gaussian forward and backward messages for the addition node in ReactiveMP.



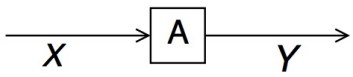
In [3]:
```
println("Forward message on Z:")
@call_rule typeof(+)(:out, Marginalisation) (m_in1 = NormalMeanVariance(1.0, 1.0), m_in2 = NormalMeanVariance(2.0, 1.
```

```
Forward message on Z:
```
Out[3]:
```
NormalMeanVariance{Float64}(μ=3.0, v=2.0)
```
In [4]:
```
println("Backward message on X:")
@call_rule typeof(+)(:in1, Marginalisation) (m_out = NormalMeanVariance(3.0, 1.0), m_in2 = NormalMeanVariance(2.0, 1.
```

```
Backward message on X:
```
Out[4]:
```
NormalMeanVariance{Float64}(μ=1.0, v=2.0)
```

Code Example: forward and backward messages for the Matrix Multiplication node

In the same way we can also investigate the forward and backward messages for the matrix multiplication ("gain") node



In [5]:
```
println("Forward message on Y:")
@call_rule typeof(*)(:out, Marginalisation) (m_A = PointMass(4.0), m_in = NormalMeanVariance(1.0, 1.0))
```

```
Forward message on Y:
```
Out[5]:`NormalMeanVariance{Float64}(μ=4.0, v=16.0)`

In [6]:
```
println("Backward message on X:")
@call_rule typeof(*)(:in, Marginalisation) (m_out = NormalMeanVariance(2.0, 1.0), m_A = PointMass(4.0), meta = TinyCo
```

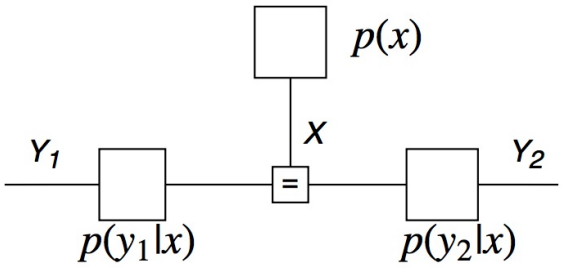
```
Backward message on X:
```
Out[6]:`NormalWeightedMeanPrecision{Float64}(xi=8.0, w=16.0)`

Example: Sum-Product Algorithm to infer a posterior

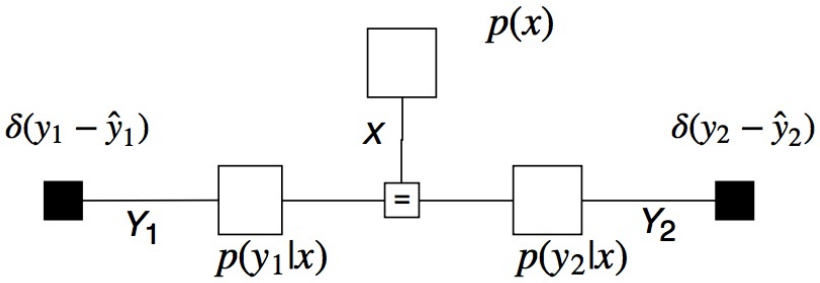
- Consider a generative model

$$p(x, y_1, y_2) = p(x)\, p(y_1|x)\, p(y_2|x).$$

  - This model expresses the assumption that y_1 and y_2 are independent measurements of x.



- Assume that we are interested in the posterior for x after observing $y_1 = \hat{y}_1$ and $y_2 = \hat{y}_2$. The posterior for x can be inferred by applying the sum-product algorithm to the following graph:



- (Note that) we usually draw terminal nodes for observed variables in the graph by smaller solid-black squares. This is just to help the visualization of the graph, since the computational rules are no different than for other nodes.

Code for Sum-Product Algorithm to infer a posterior

We'll use ReactiveMP to build the above graph, and perform sum-product message passing to infer the posterior $p(x|y_1, y_2)$. We assume $p(y_1|x)$ and $p(y_2|x)$ to be Gaussian likelihoods with known variances:

$$\begin{aligned} p(y_1 \mid x) &= \mathcal{N}(y_1 \mid x, v_{y1}) \\ p(y_2 \mid x) &= \mathcal{N}(y_2 \mid x, v_{y2}) \end{aligned}$$

Under this model, the posterior is given by:

$$\begin{aligned} p(x \mid y_1, y_2) &\propto \overbrace{p(y_1 \mid x)\, p(y_2 \mid x)}^{\text{likelihood}}\ \overbrace{p(x)}^{\text{prior}} \\ &= \mathcal{N}(x \mid \hat{y}_1, v_{y1})\, \mathcal{N}(x \mid \hat{y}_2, v_{y2})\, \mathcal{N}(x \mid m_x, v_x) \end{aligned}$$

so we can validate the answer by solving the Gaussian multiplication manually.

Continuous Data and the Gaussian Distribution

Preliminaries

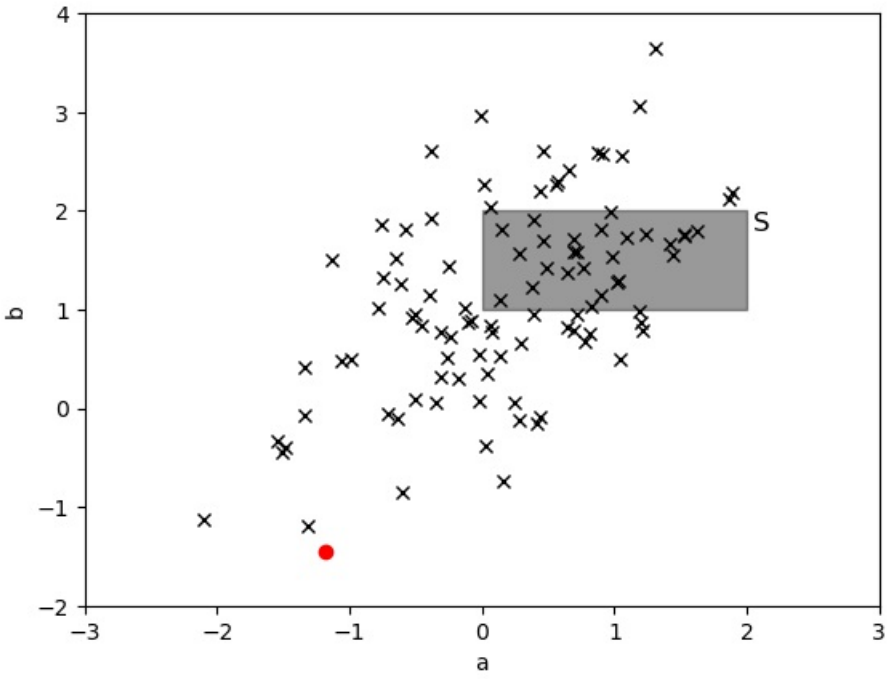
- Goal
  - Review of information processing with Gaussian distributions in linear systems
- Materials

- Mandatory
  - These lecture notes
- Optional
  - Bishop pp. 85-93
  - MacKay - 2006 - The Humble Gaussian Distribution (highly recommended!)
  - Ariel Caticha - 2012 - Entropic Inference and the Foundations of Physics, pp.30-34, section 2.8, the Gaussian distribution
- References
  - E.T. Jaynes - 2003 - Probability Theory, The Logic of Science (best book available on the Bayesian view on probability theory)

Example Problem

Consider a set of observations $D = \{x_1, \ldots, x_N\}$ in the 2-dimensional plane (see Figure). All observations were generated by the same process. We now draw an extra observation $x_\bullet = (a, b)$ from the same data generating process. What is the probability that $x_\bullet$ lies within the shaded rectangle S?

In [1]:
```
using Distributions, PyPlot
N = 100
generative_dist = MvNormal([0,1.], [0.8 0.5; 0.5 1.0])
function plotObservations(obs::Matrix)
    plot(obs[1,:], obs[2,:], "kx", zorder=3)
    fill_between([0., 2.], 1., 2., color="k", alpha=0.4, zorder=2) # Shaded area
    text(2.05, 1.8, "S", fontsize=12)
    xlim([-3,3]); ylim([-2,4]); xlabel("a"); ylabel("b")
end
D = rand(generative_dist, N) # Generate observations from generative_dist
plotObservations(D)
x_dot = rand(generative_dist) # Generate x·
plot(x_dot[1], x_dot[2], "ro");
```



The Gaussian Distribution

- Consider a random (vector) variable $x \in \mathbb{R}^M$ that is "normally" (i.e., Gaussian) distributed. The *moment* parameterization of the Gaussian distribution is completely specified by its *mean* μ and *variance* Σ and given by

$$p(x|\mu,\Sigma) = \mathcal{N}(x|\mu,\Sigma) \triangleq \frac{1}{\sqrt{(2\pi)^M |\Sigma|}} \exp\left\{-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right\}.$$

  where $|\Sigma| \triangleq \det(\Sigma)$ is the determinant of Σ.
  - For the scalar real variable $x \in \mathbb{R}$, this works out to

$$p(x|\mu,\sigma^2) = \frac{1}{\sqrt{(2\pi)\sigma^2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}.$$

- Alternatively, the _canonical_ (a.k.a. _natural_ or _information_) parameterization of the Gaussian distribution is given by

$$p(x|\eta,\Lambda) = \mathcal{N}_c(x|\eta,\Lambda) = \exp\left\{a + \eta^T x - \frac{1}{2}x^T\Lambda x\right\}.$$

- $a = -\frac{1}{2}\left(M\log(2\pi) - \log|\Lambda| + \eta^T\Lambda\eta\right)$ is the normalizing constant that ensures that $\int p(x)\mathrm{d}x = 1$.
- $\Lambda = \Sigma^{-1}$ is called the *precision matrix*.
- $\eta = \Sigma^{-1}\mu$ is the *natural* mean or for clarity often called the *precision-weighted* mean.

Why the Gaussian?

- Why is the Gaussian distribution so ubiquitously used in science and engineering? (see also Jaynes, section 7.1.4, and the whole chapter 7 in his book).

- (1) Operations on probability distributions tend to lead to Gaussian distributions:
  - Any smooth function with single rounded maximum, if raised to higher and higher powers, goes into a Gaussian function. (Example: sequential Bayesian inference).
  - The Gaussian distribution has higher entropy than any other with the same variance.
    - Therefore any operation on a probability distribution that discards information but preserves variance gets us closer to a Gaussian.
    - As an example, see Jaynes, section 7.1.4 for how this leads to the Central Limit Theorem, which results from performing convolution operations on distributions.

- (2) Once the Gaussian has been attained, this form tends to be preserved. e.g.,
  - The convolution of two Gaussian functions is another Gaussian function (useful in sum of 2 variables and linear transformations)
  - The product of two Gaussian functions is another Gaussian function (useful in Bayes rule).
  - The Fourier transform of a Gaussian function is another Gaussian function.

Transformations and Sums of Gaussian Variables

- A **linear transformation** $z = Ax + b$ of a Gaussian variable $x \sim \mathcal{N}(\mu_x, \Sigma_x)$ is Gaussian distributed as

$$p(z) = \mathcal{N}\left(z|\ A\mu_x + b, A\Sigma_x A^T\right) \tag{SRG-4a}$$

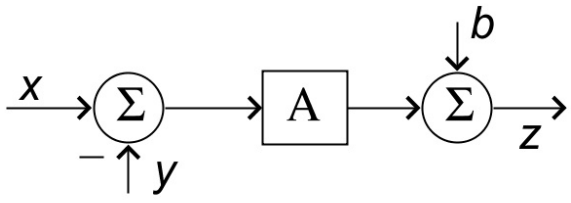
  - In fact, after a linear transformation $z = Ax + b$, no matter how x is distributed, the mean and variance of z are always given by $\mu_z = A\mu_x + b$ and $\Sigma_z = A\Sigma_x A^T$, respectively (see probability theory review lesson). In case x is not Gaussian, higher order moments may be needed to specify the distribution for z.

- The **sum of two independent Gaussian variables** is also Gaussian distributed. Specifically, if $x \sim \mathcal{N}(\mu_x, \Sigma_x)$ and $y \sim \mathcal{N}(\mu_y, \Sigma_y)$, then the PDF for $z = x + y$ is given by

$$\begin{aligned} p(z) &= \mathcal{N}(x|\ \mu_x, \Sigma_x) * \mathcal{N}(y|\ \mu_y, \Sigma_y) \\ &= \mathcal{N}\left(z\,|\,\mu_x + \mu_y, \Sigma_x + \Sigma_y\right) \end{aligned} \tag{SRG-8}$$

  - The sum of two Gaussian *distributions* is NOT a Gaussian distribution. Why not?

Example: Gaussian Signals in a Linear System



- Given independent variables $x \sim \mathcal{N}(\mu_x, \sigma_x^2)$ and $y \sim \mathcal{N}(\mu_y, \sigma_y^2)$, what is the PDF for $z = A \cdot (x - y) + b$? (see Exercises)

- Think about the role of the Gaussian distribution for stochastic linear systems in relation to what sinusoidals mean for deterministic linear system analysis.

Bayesian Inference for the Gaussian

- Let's estimate a constant θ from one 'noisy' measurement x about that constant.

- We assume the following measurement equations (the tilde $\sim$ means: 'is distributed as'):

$$\begin{aligned} x &= \theta + \epsilon \\ \epsilon &\sim \mathcal{N}(0, \sigma^2) \end{aligned}$$

- Also, let's assume a Gaussian prior for θ

$$\theta \sim \mathcal{N}(\mu_0, \sigma_0^2)$$

Model specification

- Note that you can rewrite these specifications in probabilistic notation as follows:

$$p(x|\theta) = \mathcal{N}(x|\theta, \sigma^2)$$
$$p(\theta) = \mathcal{N}(\theta|\mu_0, \sigma_0^2)$$

- (**Notational convention**). Note that we write $\epsilon \sim \mathcal{N}(0, \sigma^2)$ but not $\epsilon \sim \mathcal{N}(\epsilon|0, \sigma^2)$, and we write $p(\theta) = \mathcal{N}(\theta|\mu_0, \sigma_0^2)$ but not $p(\theta) = \mathcal{N}(\mu_0, \sigma_0^2)$.

Inference

- For simplicity, we assume that the variance σ^2 is given and will proceed to derive a Bayesian posterior for the mean θ. The case for Bayesian inference for σ^2 with a given mean is discussed in the optional slides.

- Let's do Bayes rule for the posterior PDF $p(\theta|x)$.

$$\begin{aligned}
p(\theta|x) &= \frac{p(x|\theta)p(\theta)}{p(x)} \propto p(x|\theta)p(\theta) \\
&= \mathcal{N}(x|\theta, \sigma^2)\mathcal{N}(\theta|\mu_0, \sigma_0^2) \\
&\propto \exp\left\{-\frac{(x-\theta)^2}{2\sigma^2} - \frac{(\theta-\mu_0)^2}{2\sigma_0^2}\right\} \\
&\propto \exp\left\{\theta^2 \cdot \left(-\frac{1}{2\sigma_0^2} - \frac{1}{2\sigma^2}\right) + \theta \cdot \left(\frac{\mu_0}{\sigma_0^2} + \frac{x}{\sigma^2}\right)\right\} \\
&= \exp\left\{-\frac{\sigma_0^2 + \sigma^2}{2\sigma_0^2\sigma^2}\left(\theta - \frac{\sigma_0^2 x + \sigma^2 \mu_0}{\sigma^2 + \sigma_0^2}\right)^2\right\}
\end{aligned}$$

  which we recognize as a Gaussian distribution w.r.t. θ.

- (Just as an aside,) this computational 'trick' for multiplying two Gaussians is called **completing the square**. The procedure makes use of the equality

$$ax^2 + bx + c_1 = a\left(x + \frac{b}{2a}\right)^2 + c_2$$

- In particular, it follows that the posterior for θ is

$$p(\theta|x) = \mathcal{N}(\theta|\, \mu_1, \sigma_1^2)$$

  where

$$\begin{aligned}
\frac{1}{\sigma_1^2} &= \frac{\sigma_0^2 + \sigma^2}{\sigma^2\sigma_0^2} = \frac{1}{\sigma_0^2} + \frac{1}{\sigma^2} \\
\mu_1 &= \frac{\sigma_0^2 x + \sigma^2 \mu_0}{\sigma^2 + \sigma_0^2} = \sigma_1^2\left(\frac{1}{\sigma_0^2}\mu_0 + \frac{1}{\sigma^2}x\right)
\end{aligned}$$

(Multivariate) Gaussian Multiplication

- So, multiplication of two Gaussian distributions yields another (unnormalized) Gaussian with
  - posterior precision equals **sum of prior precisions**
  - posterior precision-weighted mean equals **sum of prior precision-weighted means**

- As we just saw, great application to Bayesian inference!

$$\underbrace{\text{Gaussian}}_{\text{posterior}} \propto \underbrace{\text{Gaussian}}_{\text{likelihood}} \times \underbrace{\text{Gaussian}}_{\text{prior}}$$

- In general, the multiplication of two multi-variate Gaussians yields an (unnormalized) Gaussian:

$$\boxed{\mathcal{N}(x|\mu_a, \Sigma_a) \cdot \mathcal{N}(x|\mu_b, \Sigma_b) = \underbrace{\mathcal{N}(\mu_a|\, \mu_b, \Sigma_a + \Sigma_b)}_{\text{normalization constant}} \cdot \mathcal{N}(x|\mu_c, \Sigma_c)} \tag{SRG-6}$$

  where

$$\begin{aligned}
\Sigma_c^{-1} &= \Sigma_a^{-1} + \Sigma_b^{-1} \\
\Sigma_c^{-1}\mu_c &= \Sigma_a^{-1}\mu_a + \Sigma_b^{-1}\mu_b
\end{aligned}$$

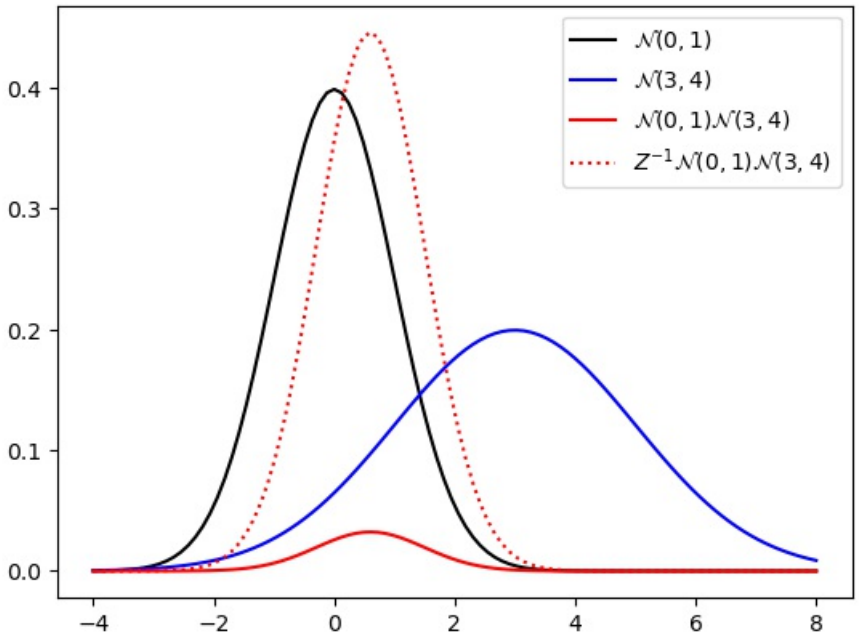
- $\Rightarrow$ Note that Bayesian inference is trivial in the *canonical* parameterization of the Gaussian, where we would get

$$\Lambda_c = \Lambda_a + \Lambda_b \quad \text{(precisions add)}$$
$$\eta_c = \eta_a + \eta_b \quad \text{(precision-weighted means add)}$$

Code Example: Product of Two Gaussian PDFs

- Let's plot the exact product of two Gaussian PDFs as well as the normalized product according to the above derivation.

In [2]:
```
using PyPlot, Distributions
d1 = Normal(0, 1) # μ=0, σ^2=1
d2 = Normal(3, 2) # μ=3, σ^2=4

# Calculate the parameters of the product d1*d2
s2_prod = (d1.σ^-2 + d2.σ^-2)^-1
m_prod = s2_prod * ((d1.σ^-2)*d1.μ + (d2.σ^-2)*d2.μ)
d_prod = Normal(m_prod, sqrt(s2_prod)) # Note that we neglect the normalization constant.

# Plot stuff
x = range(-4, stop=8, length=100)
plot(x, pdf.(d1,x), "k")
plot(x, pdf.(d2,x), "b")
plot(x, pdf.(d1,x) .* pdf.(d2,x), "r-") # Plot the exact product
plot(x, pdf.(d_prod,x), "r:")           # Plot the normalized Gaussian product
legend([L"\mathcal{N}(0,1)",
        L"\mathcal{N}(3,4)",
        L"\mathcal{N}(0,1) \mathcal{N}(3,4)",
        L"Z^{-1} \mathcal{N}(0,1) \mathcal{N}(3,4)"]);
```



The solid and dotted red curves are identical up to a scaling factor z.

Bayesian Inference with multiple Observations

- Now consider that we measure a data set $D = \{x_1, x_2, \ldots, x_N\}$, with measurements

$$x_n = \theta + \epsilon_n$$
$$\epsilon_n \sim \mathcal{N}(0, \sigma^2)$$

  and the same prior for θ:

$$\theta \sim \mathcal{N}(\mu_0, \sigma_0^2)$$

- Let's derive a distribution for the next sample x_{N+1}.

inference

- Clearly, the posterior for θ is now

$$p(\theta|D) \propto \underbrace{\mathcal{N}(\theta|\mu_0, \sigma_0^2)}_{\text{prior}} \cdot \underbrace{\prod_{n=1}^{N} \mathcal{N}(x_n|\theta, \sigma^2)}_{\text{likelihood}}$$

which is a multiplication of $N+1$ Gaussians and is therefore also Gaussian distributed.

- Using the property that precisions and precision-weighted means add when Gaussians are multiplied, we can immediately write the posterior

$$p(\theta|D) = \mathcal{N}(\theta|\,\mu_N, \sigma_N^2)$$

as

$$\frac{1}{\sigma_N^2} = \frac{1}{\sigma_0^2} + \sum_n \frac{1}{\sigma^2} \qquad \text{(B-2.142)}$$

$$\mu_N = \sigma_N^2 \left(\frac{1}{\sigma_0^2}\mu_0 + \sum_n \frac{1}{\sigma^2} x_n \right) \qquad \text{(B-2.141)}$$

application: prediction of future sample

- We now have a posterior for the model parameters. Let's write down what we know about the next sample x_{N+1}.

$$\begin{aligned} p(x_{N+1}|D_N) &= \int p(x_{N+1}|\theta) p(\theta|D_N) \mathrm{d}\theta \\ &= \int \mathcal{N}(x_{N+1}|\theta, \sigma^2) \mathcal{N}(\theta|\mu_N, \sigma_N^2) \mathrm{d}\theta \\ &= \mathcal{N}(x_{N+1}|\mu_N, \sigma_N^2 + \sigma^2) \end{aligned} \qquad \text{(use SRG-6)}$$

- Uncertainty about x_{N+1} comprises both uncertainty about the parameter (σ_N^2) and observation noise σ^2.

Maximum Likelihood Estimation for the Gaussian

- In order to determine the *maximum likelihood* estimate of θ, we let $\sigma_0^2 \to \infty$ (leads to uniform prior for θ), yielding $\frac{1}{\sigma_N^2} = \frac{N}{\sigma^2}$ and consequently

$$\mu_{\mathrm{ML}} = \mu_N|_{\sigma_0^2 \to \infty} = \sigma_N^2 \left(\frac{1}{\sigma^2} \sum_n x_n \right) = \frac{1}{N} \sum_{n=1}^N x_n$$

- As expected, having an expression for the maximum likelihood estimate, it is now possible to rewrite the (Bayesian) posterior mean for θ as (exercise)

$$\begin{aligned} \underbrace{\mu_N}_{\text{posterior}} &= \frac{\sigma^2}{N} \left(\frac{1}{\sigma_0^2}\mu_0 + \sum_n \frac{1}{\sigma^2} x_n \right) \\ &= \underbrace{\mu_0}_{\text{prior}} + \underbrace{\underbrace{\frac{N\sigma_0^2}{N\sigma_0^2 + \sigma^2}}_{\text{gain}} \cdot \underbrace{(\mu_{\mathrm{ML}} - \mu_0)}_{\text{prediction error}}}_{\text{correction}} \end{aligned} \qquad \text{(B-2.141)}$$

- Hence, the posterior mean always lies somewhere between the prior mean μ_0 and the maximum likelihood estimate (the "data" mean) μ_{ML}.

Conditioning and Marginalization of a Gaussian

- Let $z = \begin{bmatrix} x \\ y \end{bmatrix}$ be jointly normal distributed as

$$p(z) = \mathcal{N}(z|\mu, \Sigma) = \mathcal{N}\left(\begin{bmatrix} x \\ y \end{bmatrix} \middle| \begin{bmatrix} \mu_x \\ \mu_y \end{bmatrix}, \begin{bmatrix} \Sigma_x & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_y \end{bmatrix} \right)$$

- Since covariance matrices are by definition symmetric, it follows that Σ_x and Σ_y are symmetric and $\Sigma_{xy} = \Sigma_{yx}^T$.

- Let's factorize $p(z) = p(x, y)$ into $p(y|x) \cdot p(x)$ through conditioning and marginalization.

$$\text{conditioning: } \boxed{p(y|x) = \mathcal{N}\left(y|\; \mu_y + \Sigma_{yx}\Sigma_x^{-1}(x - \mu_x),\; \Sigma_y - \Sigma_{yx}\Sigma_x^{-1}\Sigma_{xy}\right)}$$

$$\text{marginalization: } \boxed{p(x) = \mathcal{N}(x|\mu_x, \Sigma_x)}$$

- **proof**: in Bishop pp.87-89

- Hence, conditioning and marginalization in Gaussians leads to Gaussians again. This is very useful for applications to Bayesian inference in jointly Gaussian systems.

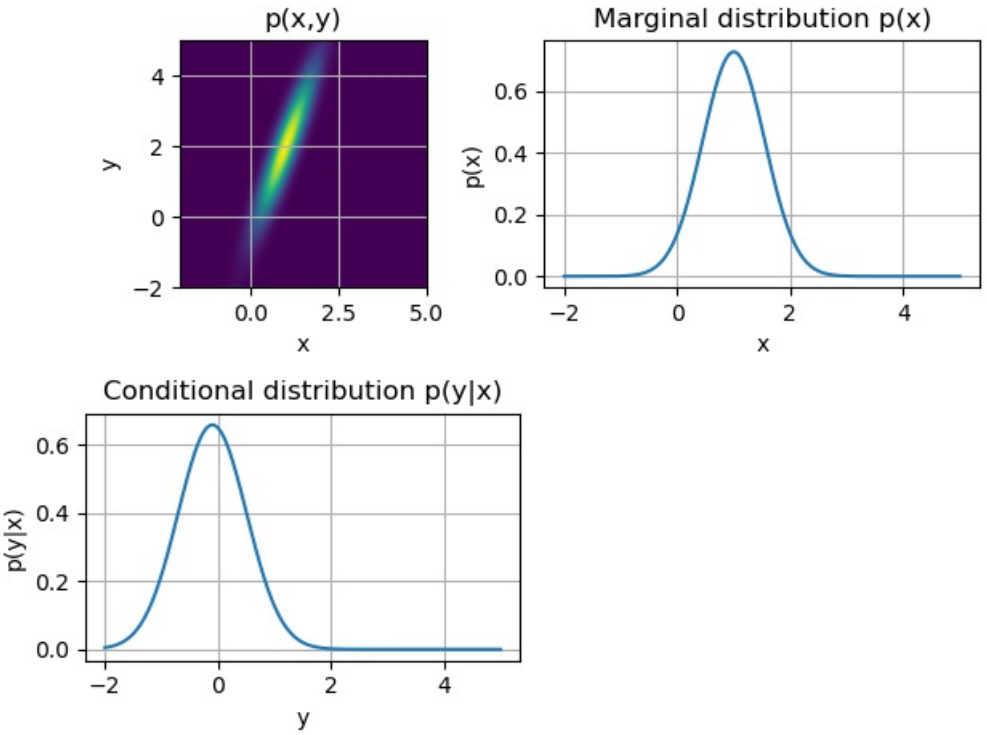
- With a natural parameterization of the Gaussian $p(z) = \mathcal{N}_c(z|\eta, \Lambda)$ with precision matrix $\Lambda = \Sigma^{-1} = \begin{bmatrix} \Lambda_x & \Lambda_{xy} \\ \Lambda_{yx} & \Lambda_y \end{bmatrix}$, the conditioning operation results in a simpler result, see Bishop pg.90, eqs. 2.96 and 2.97.

- As an exercise, interpret the formula for the conditional mean ($\mathbb{E}[y|x] = \mu_y + \Sigma_{yx}\Sigma_x^{-1}(x - \mu_x)$) as a prediction-correction operation.

Code Example: Joint, Marginal, and Conditional Gaussian Distributions

- Let's plot of the joint, marginal, and conditional distributions.

In [3]:
```
using PyPlot, Distributions

μ = [1.0; 2.0]
Σ = [0.3 0.7;
     0.7 2.0]
joint = MvNormal(μ,Σ)
marginal_x = Normal(μ[1], sqrt(Σ[1,1]))

#Plot p(x,y)
subplot(221)
x_range = y_range = range(-2,stop=5,length=1000)
joint_pdf = [ pdf(joint, [x_range[i];y_range[j]]) for  j=1:length(y_range), i=1:length(x_range)]
imshow(joint_pdf, origin="lower", extent=[x_range[1], x_range[end], y_range[1], y_range[end]])
grid(); xlabel("x"); ylabel("y"); title("p(x,y)"); tight_layout()

# Plot p(x)
subplot(222)
plot(range(-2,stop=5,length=1000), pdf.(marginal_x, range(-2,stop=5,length=1000)))
grid(); xlabel("x"); ylabel("p(x)"); title("Marginal distribution p(x)"); tight_layout()

# Plot p(y|x)
x = 0.1
conditional_y_m = μ[2]+Σ[2,1]*inv(Σ[1,1])*(x-μ[1])
conditional_y_s2 = Σ[2,2] - Σ[2,1]*inv(Σ[1,1])*Σ[1,2]
conditional_y = Normal(conditional_y_m, sqrt.(conditional_y_s2))
subplot(223)
plot(range(-2,stop=5,length=1000), pdf.(conditional_y, range(-2,stop=5,length=1000)))
grid(); xlabel("y"); ylabel("p(y|x)"); title("Conditional distribution p(y|x)"); tight_layout()
```



As is clear from the plots, the conditional distribution is a renormalized slice from the joint distribution.

Example: Conditioning of Gaussian

- Consider (again) the system

$$p(x \mid \theta) = \mathcal{N}(x \mid \theta, \sigma^2)$$

with a Gaussian prior for θ:

$$p(\theta) = \mathcal{N}(\theta \,|\, \mu_0, \sigma_0^2)$$

- Let $z = \begin{bmatrix} x \\ \theta \end{bmatrix}$. The distribution for z is then given by (Exercise)

$$p(z) = p\left(\begin{bmatrix} x \\ \theta \end{bmatrix}\right) = \mathcal{N}\left(\begin{bmatrix} x \\ \theta \end{bmatrix} \middle| \begin{bmatrix} \mu_0 \\ \mu_0 \end{bmatrix}, \begin{bmatrix} \sigma_0^2 + \sigma^2 & \sigma_0^2 \\ \sigma_0^2 & \sigma_0^2 \end{bmatrix}\right)$$

- Direct substitution of the rule for Gaussian conditioning leads to the posterior (derivation as an Exercise):

$$p(\theta|x) = \mathcal{N}\left(\theta|\, \mu_1, \sigma_1^2\right),$$

with

$$\begin{aligned} K &= \frac{\sigma_0^2}{\sigma_0^2 + \sigma^2} \qquad (K \text{ is called: Kalman gain}) \\ \mu_1 &= \mu_0 + K \cdot (x - \mu_0) \\ \sigma_1^2 &= (1 - K)\, \sigma_0^2 \end{aligned}$$

- $\Rightarrow$ Moral: For jointly Gaussian systems, we can do inference simply in one step by using the formulas for conditioning and marginalization.

Recursive Bayesian Estimation

- Consider the signal $x_t = \theta + \epsilon_t$, where $D_t = \{x_1, \ldots, x_t\}$ is observed *sequentially* (over time).
- **Problem**: Derive a recursive algorithm for $p(\theta|D_t)$, i.e., an update rule for (posterior) $p(\theta|D_t)$ based on (prior) $p(\theta|D_{t-1})$ and (new observation) x_t.

Model specification

- Let's define the estimate after t observations (i.e., our *solution*) as $p(\theta|D_t) = \mathcal{N}(\theta|\, \mu_t, \sigma_t^2)$.
- We define the joint distribution for θ and x_t, given background D_{t-1}, by

$$\begin{aligned} p(x_t, \theta|\, D_{t-1}) &= p(x_t|\theta)\, p(\theta|D_{t-1}) \\ &= \underbrace{\mathcal{N}(x_t \,|\, \theta, \sigma^2)}_{\text{likelihood}} \underbrace{\mathcal{N}(\theta|\, \mu_{t-1}, \sigma_{t-1}^2)}_{\text{prior}} \end{aligned}$$

Inference

- Use Bayes rule,

$$\begin{aligned} p(\theta|D_t) &= p(\theta|x_t, D_{t-1}) \\ &\propto p(x_t, \theta|D_{t-1}) \\ &= p(x_t|\theta)\, p(\theta|D_{t-1}) \\ &= \mathcal{N}(x_t|\theta, \sigma^2)\, \mathcal{N}(\theta|\, \mu_{t-1}, \sigma_{t-1}^2) \\ &= \mathcal{N}(\theta|x_t, \sigma^2)\, \mathcal{N}(\theta|\, \mu_{t-1}, \sigma_{t-1}^2) \quad (\text{note this trick}) \\ &= \mathcal{N}(\theta|\mu_t, \sigma_t^2) \quad (\text{use Gaussian multiplication formula SRG-6}) \end{aligned}$$

with

$$\begin{aligned} K_t &= \frac{\sigma_{t-1}^2}{\sigma_{t-1}^2 + \sigma^2} \qquad (\text{Kalman gain}) \\ \mu_t &= \mu_{t-1} + K_t \cdot (x_t - \mu_{t-1}) \\ \sigma_t^2 &= (1 - K_t)\, \sigma_{t-1}^2 \end{aligned}$$

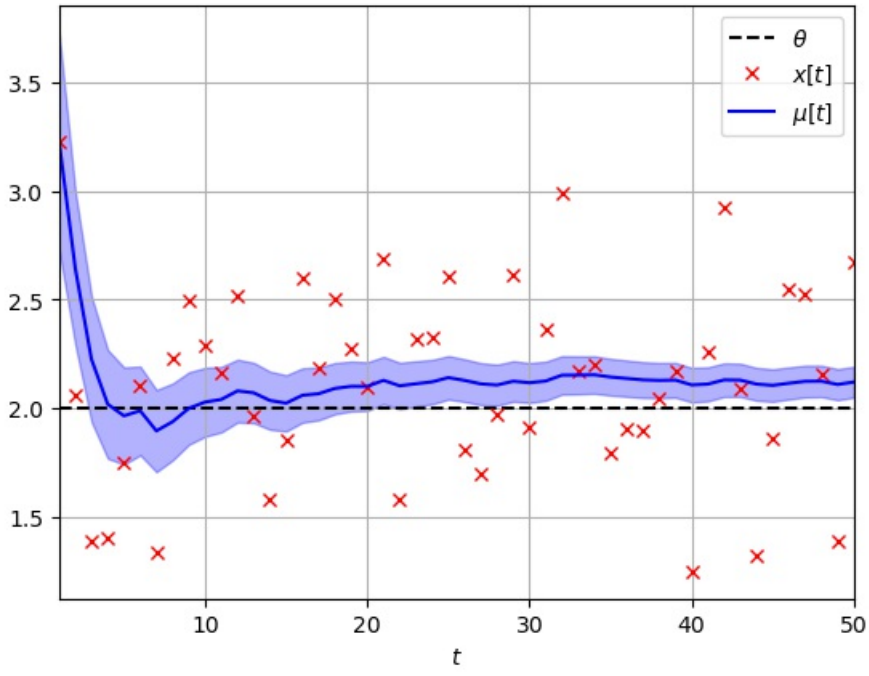
- This linear *sequential* estimator of mean and variance in Gaussian observations is called a **Kalman Filter**.
- Note that the uncertainty about θ decreases over time (since $0 < (1 - K_t) < 1$). Since we assume that the statistics of the system do not change (stationarity), each new sample provides new information.
- Recursive Bayesian estimation is the basis for **adaptive signal processing** algorithms such as Least Mean Squares (LMS) and Recursive Least Squares (RLS).

Code Example: Kalman Filter

- Let's implement the Kalman filter described above. We'll use it to recursively estimate the value of θ based on noisy observations.

In [4]:
```
using PyPlot

N = 50                          # Number of observations
```

```
θ = 2.0                         # True value of the variable we want to estimate
σ_ϵ2 = 0.25                     # Observation noise variance
x = sqrt(σ_ϵ2) * randn(N) .+ θ # Generate N noisy observations of θ

t = 0
μ = fill!(Vector{Float64}(undef,N), NaN)     # Means of p(θ|D) over time
σ_μ2 = fill!(Vector{Float64}(undef,N), NaN) # Variances of p(θ|D) over time

function performKalmanStep()
    # Perform a Kalman filter step, update t, μ, σ_μ2
    global t += 1
    if t>1 # Use posterior from prev. step as prior
        K = σ_μ2[t-1] / (σ_ϵ2 + σ_μ2[t-1]) # Kalman gain
        μ[t] = μ[t-1] + K*(x[t] - μ[t-1])  # Update mean using (1)
        σ_μ2[t] = σ_μ2[t-1] * (1.0-K)      # Update variance using (2)
    elseif t==1 # Use prior
        # Prior p(θ) = N(0,1000)
        K = 1000.0 / (σ_ϵ2 + 1000.0) # Kalman gain
        μ[t] = 0 + K*(x[t] - 0)      # Update mean using (1)
        σ_μ2[t] = 1000 * (1.0-K)     # Update variance using (2)
    end
end

while t<N
    performKalmanStep()
end

# Plot the 'true' value of θ, noisy observations x, and the recursively updated posterior p(θ|D)
t = collect(1:N)
plot(t, θ*ones(N), "k--")
plot(t, x, "rx")
plot(t, μ, "b-")
fill_between(t, μ-sqrt.(σ_μ2), μ+sqrt.(σ_μ2), color="b", alpha=0.3)
legend([L"\theta", L"x[t]", L"\mu[t]"])
xlim((1, N)); xlabel(L"t"); grid()
```



The shaded area represents 2 standard deviations of posterior $p(\theta|D)$. The variance of the posterior is guaranteed to decrease monotonically for the standard Kalman filter.

Product of Normally Distributed Variables

- (We've seen that) the sum of two Gausssian distributed variables is also Gaussian distributed.
- Has the *product* of two Gaussian distributed variables also a Gaussian distribution?
- **No**! In general this is a difficult computation. As an example, let's compute $p(z)$ for $Z = XY$ for the special case that $X \sim \mathcal{N}(0,1)$ and $Y \sim \mathcal{N}(0,1)$.

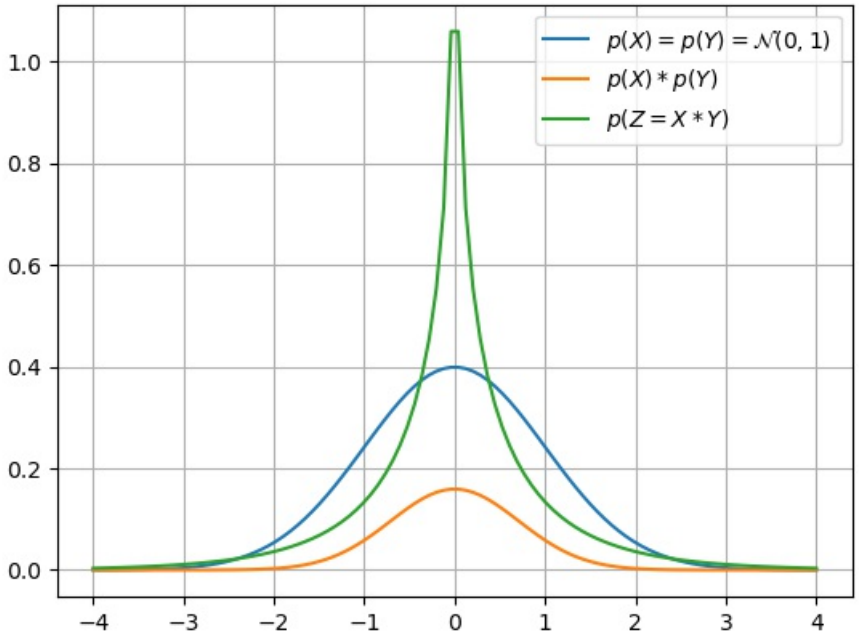
$$
\begin{aligned}
p(z) &= \int_{X,Y} p(z|x,y)\, p(x,y)\, \mathrm{d}x\mathrm{d}y \\
&= \frac{1}{2\pi} \int \delta(z - xy)\, e^{-(x^2+y^2)/2}\, \mathrm{d}x\mathrm{d}y \\
&= \frac{1}{\pi} \int_0^\infty \frac{1}{x} e^{-(x^2+z^2/x^2)/2}\, \mathrm{d}x \\
&= \frac{1}{\pi} \mathrm{K}_0(|z|)\,.
\end{aligned}
$$

where $\mathrm{K}_n(z)$ is a modified Bessel function of the second kind.

Code Example: Product of Gaussian Distributions

- We plot $p(Z = XY)$ and $p(X)p(Y)$ for $X \sim \mathcal{N}(0,1)$ and $Y \sim \mathcal{N}(0,1)$ to give an idea of how these distributions differ.

In [5]:
```
using PyPlot, Distributions, SpecialFunctions
X = Normal(0,1)
Y = Normal(0,1)
pdf_product_std_normals(z::Vector) = (besselk.(0, abs.(z))./π)
range1 = collect(range(-4,stop=4,length=100))
plot(range1, pdf.(X, range1))
plot(range1, pdf.(X,range1).*pdf.(Y,range1))
plot(range1, pdf_product_std_normals(range1))
legend([L"p(X)=p(Y)=\mathcal{N}(0,1)", L"p(X)*p(Y)",L"p(Z=X*Y)"]); grid()
```



- In short, Gaussian-distributed variables remain Gaussian in linear systems, but this is not the case in non-linear systems.

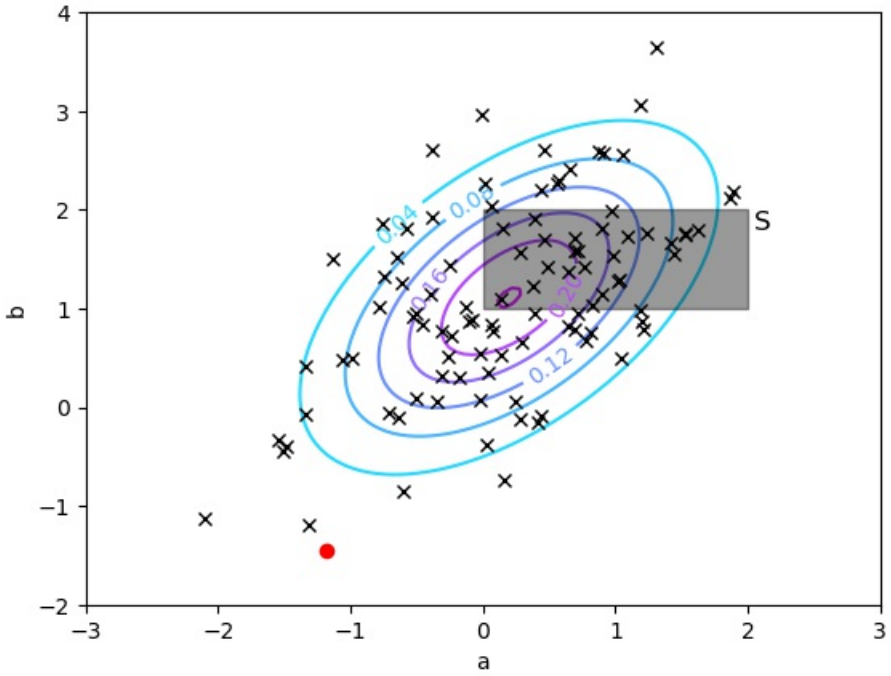
Solution to Example Problem

We apply maximum likelihood estimation to fit a 2-dimensional Gaussian model (m) to data set D. Next, we evaluate $p(x_\bullet \in S|m)$ by (numerical) integration of the Gaussian pdf over S: $p(x_\bullet \in S|m) = \int_S p(x|m)\mathrm{d}x$.

In [6]:
```
using HCubature, LinearAlgebra# Numerical integration package
# Maximum likelihood estimation of 2D Gaussian
N = length(sum(D,dims=1))
μ = 1/N * sum(D,dims=2)[:,1]
D_min_μ = D - repeat(μ, 1, N)
Σ = Hermitian(1/N * D_min_μ*D_min_μ')
m = MvNormal(μ, convert(Matrix, Σ));

# # Contour plot of estimated Gaussian density
A = Matrix{Float64}(undef,100,100); B = Matrix{Float64}(undef,100,100)
density = Matrix{Float64}(undef,100,100)
for i=1:100
    for j=1:100
        A[i,j] = a = (i-1)*6/100 .- 2
        B[i,j] = b = (j-1)*6/100 .- 3
        density[i,j] = pdf(m, [a,b])
    end
end
c = contour(A, B, density, 6, zorder=1)
```

```
    PyPlot.set_cmap("cool")
    clabel(c, inline=1, fontsize=10)

    # Plot observations, x·, and the countours of the estimated Gausian density
    plotObservations(D)
    plot(x_dot[1], x_dot[2], "ro")

    # Numerical integration of p(x|m) over S:
    (val,err) = hcubature((x)->pdf(m,x), [0., 1.], [2., 2.])
    println("p(x·∈S|m) ≈ $(val)")
```

```
p(x·∈S|m) ≈ 0.25614698349481774
```



Summary

- A linear transformation of a Gaussian-distributed (potentially multivariate) variable remains Gaussian.
- Bayesian inference with Gaussian prior and likelihood leads to an analytically computable Gaussian posterior.
- Here's a nice summary of Gaussian calculations by Sam Roweis.

OPTIONAL SLIDES

Inference for the Precision Parameter of the Gaussian

- Again, we consider an observed data set $D = \{x_1, x_2, \ldots, x_N\}$ and try to explain these data by a Gaussian distribution.
- We discussed earlier Bayesian inference for the mean with a given variance. Now we will derive a posterior for the variance if the mean is given. (Technically, we will do the derivation for a precision parameter $\lambda = \sigma^{-2}$, since the discussion is a bit more straightforward for the precision parameter).

model specification

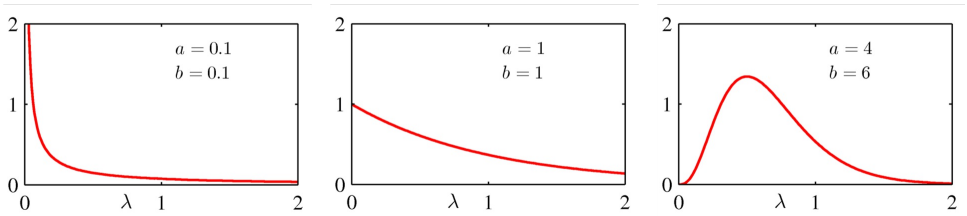
- The likelihood for the precision parameter is

$$\begin{aligned} p(D|\lambda) &= \prod_{n=1}^{N} \mathcal{N}\left(x_n \,|\, \mu, \lambda^{-1}\right) \\ &\propto \lambda^{N/2} \exp\left\{-\frac{\lambda}{2}\sum_{n=1}^{N}(x_n-\mu)^2\right\} \end{aligned} \tag{B-2.145}$$

- The conjugate distribution for this function of λ is the *Gamma* distribution, given by

$$p(\lambda|\, a, b) = \mathrm{Gam}\,(\lambda|\, a, b) \triangleq \frac{1}{\Gamma(a)} b^a \lambda^{a-1} \exp\{-b\lambda\}, \tag{B-2.146}$$

where $a > 0$ and $b > 0$ are known as the *shape* and *rate* parameters, respectively.



- (Bishop fig.2.13). Plots of the Gamma distribution $\text{Gam}(\lambda \,|\, a, b)$ for different values of a and b.
- The mean and variance of the Gamma distribution evaluate to $\text{E}(\lambda) = \frac{a}{b}$ and $\text{var}\,[\lambda] = \frac{a}{b^2}$.

inference

- We will consider a prior $p(\lambda) = \text{Gam}(\lambda \,|\, a_0, b_0)$, which leads by Bayes rule to the posterior

$$p(\lambda \,|\, D) \propto \underbrace{\lambda^{N/2} \exp\left\{ -\frac{\lambda}{2} \sum_{n=1}^{N} (x_n - \mu)^2 \right\}}_{\text{likelihood}} \cdot \underbrace{\frac{1}{\Gamma(a_0)} b_0^{a_0} \lambda^{a_0 - 1} \exp\{-b_0 \lambda\}}_{\text{prior}} \tag{1}$$

$$\propto \text{Gam}\,(\lambda \,|\, a_N, b_N) \tag{2}$$

with

$$a_N = a_0 + \frac{N}{2} \tag{B-2.150}$$

$$b_N = b_0 + \frac{1}{2} \sum_n (x_n - \mu)^2 \tag{B-2.151}$$

- Hence the **posterior is again a Gamma distribution**. By inspection of B-2.150 and B-2.151, we deduce that we can interpret $2a_0$ as the number of a priori (pseudo-)observations.
- Since the most uninformative prior is given by $a_0 = b_0 \to 0$, we can derive the **maximum likelihood estimate** for the precision as

$$\lambda_{\text{ML}} = \text{E}\,[\lambda]|_{a_0 = b_0 \to 0} = \left. \frac{a_N}{b_N} \right|_{a_0 = b_0 \to 0} = \frac{N}{\sum_{n=1}^{N} (x_n - \mu)^2}$$

Discrete Data and the Multinomial Distribution

Preliminaries

- Goal
  - Simple Bayesian and maximum likelihood-based density estimation for discretely valued data sets
- Materials
  - Mandatory
    - These lecture notes
  - Optional
    - Bishop pp. 67-70, 74-76, 93-94

Discrete Data: the 1-of-K Coding Scheme

- Consider a coin-tossing experiment with outcomes $x \in \{0, 1\}$ (tail and head) and let $0 \leq \mu \leq 1$ represent the probability of heads. This model can written as a **Bernoulli distribution**:

$$p(x|\mu) = \mu^x (1 - \mu)^{1 - x}$$

  - Note that the variable x acts as a (binary) **selector** for the tail or head probabilities. Think of this as an 'if'-statement in programming.
- Now consider a K-sided coin (e.g., a six-faced *die* (pl.: dice)). How should we encode outcomes?
- **Option 1**: $x \in \{1, 2, \ldots, K\}$.
  - E.g., for $K = 6$, if the die lands on the 3rd face $\Rightarrow x = 3$.
- **Option 2**: $x = (x_1, \ldots, x_K)^T$ with **binary selection variables**

$$x_k = \begin{cases} 1 & \text{if die landed on } k\text{th face} \\ 0 & \text{otherwise} \end{cases}$$

- E.g., for $K=6$, if the die lands on the 3rd face $\Rightarrow x=(0,0,1,0,0,0)^T$.
- This coding scheme is called a **1-of-K** or **one-hot** coding scheme.

- It turns out that the one-hot coding scheme is mathematically more convenient!

The Categorical Distribution

- Consider a K-sided die. We use a one-hot coding scheme. Assume the probabilities $p(x_k=1)=\mu_k$ with $\sum_k \mu_k = 1$. The data generating distribution is then (note the similarity to the Bernoulli distribution)

$$p(x|\mu) = \mu_1^{x_1}\mu_2^{x_2}\cdots\mu_K^{x_K} = \prod_{k=1}^{K}\mu_k^{x_k} \tag{B-2.26}$$

- This generalized Bernoulli distribution is called the **categorical distribution** (or sometimes the 'multi-noulli' distribution).

Bayesian Density Estimation for a Loaded Die

- Now let's proceed with Bayesian density estimation, i.e., let's learn the parameters for model $p(x|\theta)$ for an observed data set $D=\{x_1,\ldots,x_N\}$ of N IID rolls of a K-sided die, with

$$x_{nk} = \begin{cases} 1 & \text{if the } n\text{th throw landed on } k\text{th face} \\ 0 & \text{otherwise} \end{cases}$$

Model specification

- The data generating PDF is

$$p(D|\mu) = \prod_n\prod_k \mu_k^{x_{nk}} = \prod_k \mu_k^{\sum_n x_{nk}} = \prod_k \mu_k^{m_k} \tag{B-2.29}$$

  where $m_k = \sum_n x_{nk}$ is the total number of occurrences that we 'threw' k eyes. Note that $\sum_k m_k = N$.
  - This distribution depends on the observations **only** through the quantities $\{m_k\}$.

- We need a prior for the parameters $\mu=(\mu_1,\mu_2,\ldots,\mu_K)$. In the binary coin toss example, we used a beta distribution that was conjugate with the binomial and forced us to choose prior pseudo-counts.

- The generalization of the beta prior to the K parameters $\{\mu_k\}$ is the Dirichlet distribution:

$$p(\mu|\alpha) = \mathrm{Dir}(\mu|\alpha) = \frac{\Gamma\left(\sum_k \alpha_k\right)}{\Gamma(\alpha_1)\cdots\Gamma(\alpha_K)}\prod_{k=1}^{K}\mu_k^{\alpha_k-1}$$

  where $\Gamma(\cdot)$ is the Gamma function.

- The Gamma function can be interpreted as a generalization of the factorial function to the real ($\mathbb{R}$) numbers. If n is a natural number ($1,2,3,\ldots$), then $\Gamma(n)=(n-1)!$, where $(n-1)!=(n-1)\cdot(n-2)\cdot 1$.

- As before for the Beta distribution in the coin toss experiment, you can interpret α_k-1 as the prior number of (pseudo-)observations that the die landed on the kth face.

Inference for $\{\mu_k\}$

- The posterior for $\{\mu_k\}$ can be obtained through Bayes rule:

$$\begin{aligned} p(\mu|D,\alpha) &\propto p(D|\mu)\cdot p(\mu|\alpha) \\ &\propto \prod_k \mu_k^{m_k}\cdot\prod_k \mu_k^{\alpha_k-1} \\ &= \prod_k \mu_k^{\alpha_k+m_k-1} \\ &\propto \mathrm{Dir}\left(\mu|\,\alpha+m\right) \\ &= \frac{\Gamma\left(\sum_k(\alpha_k+m_k)\right)}{\Gamma(\alpha_1+m_1)\Gamma(\alpha_2+m_2)\cdots\Gamma(\alpha_K+m_K)}\prod_{k=1}^{K}\mu_k^{\alpha_k+m_k-1} \end{aligned} \tag{B-2.41}$$

where $m=(m_1,m_2,\ldots,m_K)^T$.

- We recognize the (α_k-1)'s as prior pseudo-counts and the Dirichlet distribution shows to be a conjugate prior to the categorical/multinomial:

$$\underbrace{\mathrm{Dirichlet}}_{\text{posterior}} \propto \underbrace{\mathrm{categorical}}_{\text{likelihood}}\cdot\underbrace{\mathrm{Dirichlet}}_{\text{prior}}$$

- This is actually a generalization of the conjugate relation that we found for the binary coin toss:

$$\underbrace{\text{Beta}}_{\text{posterior}} \propto \underbrace{\text{binomial}}_{\text{likelihood}} \cdot \underbrace{\text{Beta}}_{\text{prior}}$$

Prediction of next toss for the loaded die

- Let's apply what we have learned about the loaded die to compute the probability that we throw the k-th face at the next toss.

$$\begin{aligned} p(x_{\bullet,k}=1|D) &= \int p(x_{\bullet,k}=1|\mu)\, p(\mu|D)\, \mathrm{d}\mu \\ &= \int_0^1 \mu_k \times \mathcal{D}ir(\mu|\,\alpha+m)\, \mathrm{d}\mu \\ &= \mathrm{E}\left[\mu_k\right] \\ &= \frac{m_k+\alpha_k}{N+\sum_k \alpha_k} \end{aligned}$$

- (You can find the mean of the Dirichlet distribution at its Wikipedia site).
- This result is simply a generalization of **Laplace's rule of succession**.

Categorical, Multinomial and Related Distributions

- In the above derivation, we noticed that the data generating distribution for N die tosses $D=\{x_1,\ldots,x_N\}$ only depends on the **data frequencies** m_k:

$$p(D|\mu) = \prod_n \underbrace{\prod_k \mu_k^{x_{nk}}}_{\text{categorical dist.}} = \prod_k \mu_k^{\sum_n x_{nk}} = \prod_k \mu_k^{m_k} \tag{B-2.29}$$

- A related distribution is the distribution over data frequency observations $D_m=\{m_1,\ldots,m_K\}$, which is called the **multinomial distribution**,

$$p(D_m|\mu) = \frac{N!}{m_1!m_2!\ldots m_K!}\prod_k \mu_k^{m_k}.$$

- When used as a likelihood function for μ, it makes no difference whether you use $p(D|\mu)$ or $p(D_m|\mu)$. Why?

- Verify for yourself that (Exercise):
  - the categorial distribution is a special case of the multinomial for $N=1$.
  - the Bernoulli is a special case of the categorial distribution for $K=2$.
  - the binomial is a special case of the multinomial for $K=2$.

Maximum Likelihood Estimation for the Multinomial

Maximum likelihood as a special case of Bayesian estimation

- We can get the maximum likelihood estimate $\hat{\mu}_k$for μ_k based on N throws of a K-sided die from the Bayesian framework by using a uniform prior for μ and taking the mode of the posterior for μ:

$$\begin{aligned} \hat{\mu}_k &= \arg\max_{\mu_k} p(D|\mu) \\ &= \arg\max_{\mu_k} p(D|\mu)\cdot \text{Uniform}(\mu) \\ &= \arg\max_{\mu_k} p(D|\mu)\cdot \text{Dir}(\mu|\alpha)|_{\alpha=(1,1,\ldots,1)} \\ &= \arg\max_{\mu_k} p(\mu|D,\alpha)|_{\alpha=(1,1,\ldots,1)} \\ &= \arg\max_{\mu_k} \text{Dir}\,(\mu|m+\alpha)|_{\alpha=(1,1,\ldots,1)} \\ &= \frac{m_k}{\sum_k m_k} = \frac{m_k}{N} \end{aligned}$$

  where we used the fact that the maximum of the Dirichlet distribution $\text{Dir}(\{\alpha_1,\ldots,\alpha_K\})$ is obtained at $(\alpha_k-1)/(\sum_k \alpha_k - K)$.

Maximum likelihood estimation by optimizing a constrained log-likelihood

- Of course, we shouldn't have to go through the full Bayesian framework to get the maximum likelihood estimate. Alternatively, we can find the maximum of the likelihood directly.
- The log-likelihood for the multinomial distribution is given by

$$\mathrm{L}(\mu) \triangleq \log p(D_m|\mu) \propto \log \prod_k \mu_k^{m_k} = \sum_k m_k \log \mu_k$$

- When doing ML estimation, we must obey the constraint $\sum_k \mu_k = 1$, which can be accomplished by a Lagrange multiplier (see Bishop App.E). The **augmented log-likelihood** with Lagrange multiplier is then

$$\mathrm{L}'(\mu) = \sum_k m_k \log \mu_k + \lambda\cdot\left(1-\sum_k \mu_k\right)$$

- Set derivative to zero yields the **sample proportion** for μ_k

$$\nabla_{\mu_k} \mathrm{L}' = \frac{m_k}{\mu_k} - \lambda \stackrel{!}{=} 0 \;\Rightarrow\; \mu_k = \frac{m_k}{N}$$

where we get λ from the constraint

$$\sum_k \mu_k = \sum_k \frac{m_k}{\lambda} = \frac{N}{\lambda} \stackrel{!}{=} 1$$

Recap Maximum Likelihood Estimation for Gaussian and Multinomial Distributions

Given N IID observations $D = \{x_1, \ldots, x_N\}$.

- For a **multivariate Gaussian** model $p(x_n|\theta) = \mathcal{N}(x_n|\mu, \Sigma)$, we obtain ML estimates

$$\hat{\mu} = \frac{1}{N}\sum_n x_n \qquad \text{(sample mean)}$$
$$\hat{\Sigma} = \frac{1}{N}\sum_n (x_n - \hat{\mu})(x_n - \hat{\mu})^T \qquad \text{(sample variance)}$$

- For discrete outcomes modeled by a 1-of-K **categorical distribution** we find

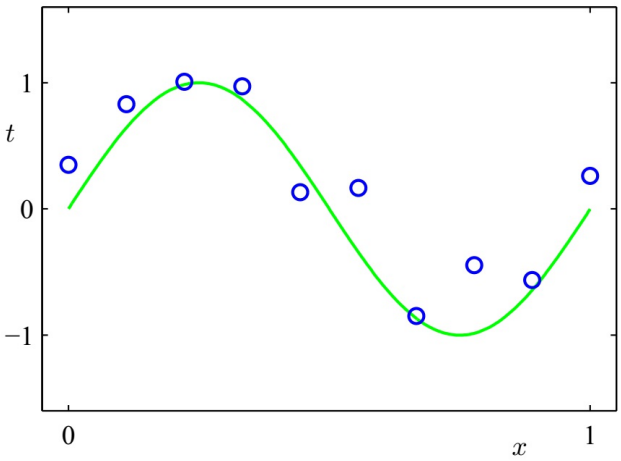
$$\hat{\mu}_k = \frac{1}{N}\sum_n x_{nk} \quad \left(= \frac{m_k}{N}\right) \qquad \text{(sample proportion)}$$

Regression

Preliminaries

- Goal
  - Introduction to Bayesian (Linear) Regression
- Materials
  - Mandatory
    - These lecture notes
  - Optional
    - Bishop pp. 152-158
    - In this and forthcoming lectures, we will make use of some elementary matrix calculus. The most important formulas are summarized at the bottom of this notebook in an OPTIONAL SLIDE on matrix calculus. For derivations, see Bishop Appendix C.

Regression - Illustration



Given a set of (noisy) data measurements, find the 'best' relation between an input variable $x \in \mathbb{R}^D$ and input-dependent outcomes $y \in \mathbb{R}$.

Regression vs Density Estimation

- Observe N IID data **pairs** $D = \{(x_1, y_1), \ldots, (x_N, y_N)\}$ with $x_n \in \mathbb{R}^M$ and $y_n \in \mathbb{R}$.

- Assume that we are interested in (a model for) the responses y_n for **given inputs** x_n?, I.o.w. we are interested in building a model for the conditional distribution $p(y|x)$.

- Note that, since $p(x,y) = p(y|x)\,p(x)$, building a model $p(y|x)$ is similar to density estimation with the assumption that x is drawn from a uniform distribution.

Bayesian Linear Regression

- Next, we discuss (1) model specification, (2) Inference and (3) a prediction application for a Bayesian linear regression problem.

1. Model Specification

- In a traditional *regression* model, we try to 'explain the data' by a purely deterministic function $f(x_n, w)$, plus a purely random term ϵ_n for 'unexplained noise':

$$y_n = f(x_n, w) + \epsilon_n$$

- In a *linear regression* model, i.e., linear w.r.t. the parameters w, we assume that

$$f(x_n, w) = \sum_{j=0}^{M-1} w_j \phi_j(x_n) = w^T \phi(x_n)$$

  where $\phi_j(x)$ are called basis functions.
  - For notational simplicity, from now on we will assume $f(x_n, w) = w^T x_n$, with $x_n \in \mathbb{R}^M$.

- In *ordinary linear regression* , the noise process ϵ_n is zero-mean Gaussian with constant variance, i.e.

$$\epsilon_n \sim \mathcal{N}(0, \beta^{-1})\,.$$

- Hence, given a data set $D = \{(x_1, y_1), \ldots, (x_N, y_N)\}$, the likelihood for an ordinary linear regression model is

$$\begin{aligned} p(y|\,\mathbf{X}, w, \beta) &= \mathcal{N}(y\,|\,\mathbf{X}w,\, \beta^{-1}\mathbf{I}) \\ &= \prod_n \mathcal{N}(y_n\,|\,w^T x_n,\, \beta^{-1}) \end{aligned} \tag{B-3.10}$$

where $w = \begin{pmatrix} w_1 \\ w_2 \\ \vdots \\ w_M \end{pmatrix}$, the $(N \times M)$-dim matrix $\mathbf{X} = \begin{pmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_N^T \end{pmatrix} = \begin{pmatrix} x_{11}, x_{12}, \ldots, x_{1M} \\ x_{21}, x_{22}, \ldots, x_{2M} \\ \vdots \\ x_{N1}, x_{N2}, \ldots, x_{NM} \end{pmatrix}$ and $y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix}$.

- For full Bayesian learning we should also choose a prior $p(w)$:

$$p(w|\,\alpha) = \mathcal{N}(w\,|\,0,\, \alpha^{-1}\mathbf{I}) \tag{B-3.52}$$

- For simplicity, we will assume that α and β are fixed and known.

2. Inference

- We'll do Bayesian inference for the parameters w.

$$\begin{aligned} p(w|D) &\propto p(D|w) \cdot p(w) \\ &= \mathcal{N}(y\,|\,\mathbf{X}w,\, \beta^{-1}\mathbf{I}) \cdot \mathcal{N}(w\,|\,0,\, \alpha^{-1}\mathbf{I}) \\ &\propto \exp\Big(-\frac{\beta}{2}(y - \mathbf{X}w)^T(y - \mathbf{X}w) - \frac{\alpha}{2}w^T w\Big) \qquad \text{(B-3.55)} \\ &= \exp\Big(-\frac{1}{2}w^T\underbrace{(\beta\mathbf{X}^T\mathbf{X} + \alpha\mathbf{I})}_{\Lambda_N}w + \underbrace{(\beta\mathbf{X}^T y)}_{\eta_N}{}^T w - \frac{\beta}{2}y^T y\Big) \\ &\propto \mathcal{N}_c(w\,|\,\eta_N, \Lambda_N) \end{aligned}$$

with natural parameters (see the natural parameterization of Gaussian):

$$\begin{aligned} \eta_N &= \beta\mathbf{X}^T y \\ \Lambda_N &= \beta\mathbf{X}^T\mathbf{X} + \alpha\mathbf{I} \end{aligned}$$

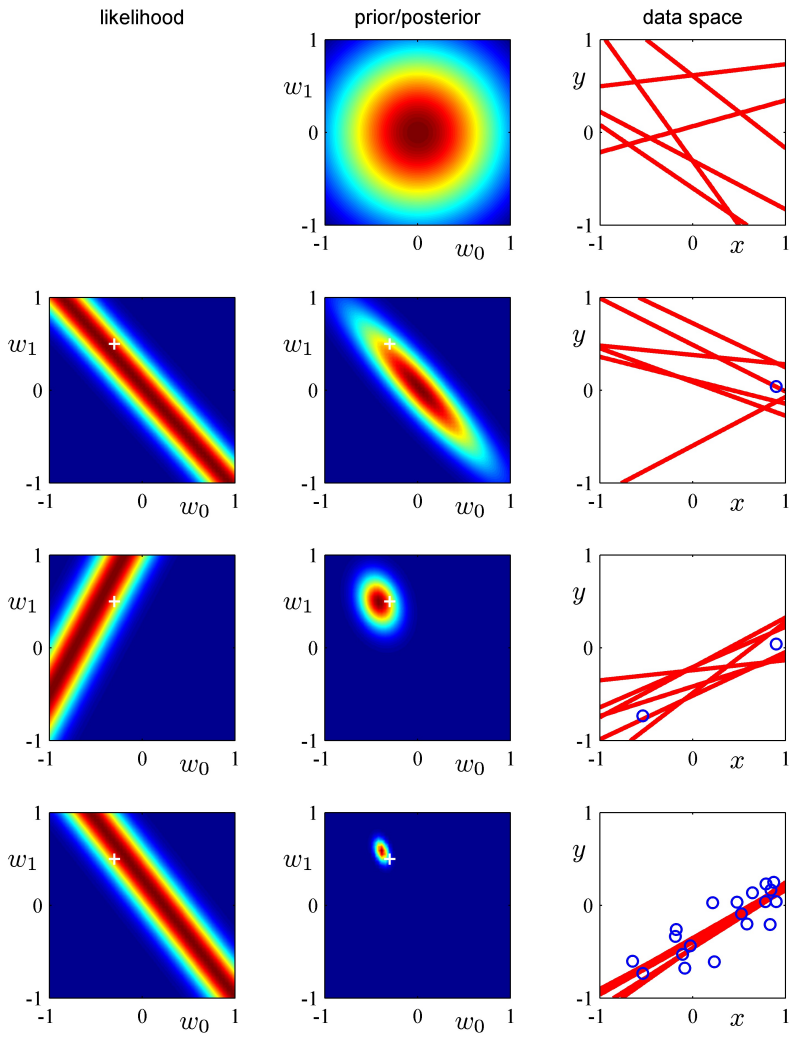
- Or equivalently (in the moment parameterization of the Gaussian):

$$\begin{aligned} p(w|D) &= \mathcal{N}(w\,|\,m_N, S_N) && \text{(B-3.49)} \\ m_N &= \beta S_N \mathbf{X}^T y && \text{(B-3.53)} \\ S_N &= \big(\alpha\mathbf{I} + \beta\mathbf{X}^T\mathbf{X}\big)^{-1} && \text{(B-3.54)} \end{aligned}$$

- Note that B-3.53 and B-3.54 combine to

$$m_N = \Big(\frac{\alpha}{\beta}\mathbf{I} + \mathbf{X}^T\mathbf{X}\Big)^{-1}\mathbf{X}^T y.$$

- (Bishop Fig.3.7) Illustration of sequential Bayesian learning for a simple linear model of the form $y(x, w) = w_0 + w_1 x$. (Bishop Fig.3.7, detailed description at Bishop, pg.154.)



3. Application: predictive distribution

- Assume we are interested in the distribution $p(y_\bullet \mid x_\bullet, D)$ for a new input $x_\bullet$. This can be worked out as (exercise B-3.10)

$$
\begin{aligned}
p(y_\bullet \mid x_\bullet, D) &= \int p(y_\bullet \mid x_\bullet, w) p(w \mid D)\, \mathrm{d}w \\
&= \int \mathcal{N}(y_\bullet \mid w^T x_\bullet, \beta^{-1}) \mathcal{N}(w \mid m_N, S_N)\, \mathrm{d}w \\
&= \int \mathcal{N}(y_\bullet \mid z, \beta^{-1}) \underbrace{\mathcal{N}(z \mid x_\bullet^T m_N, x_\bullet^T S_N x_\bullet)\, \mathrm{d}z}_{=\mathcal{N}(w \mid m_N, S_N)\, \mathrm{d}w} \quad (\text{sub. } z = x_\bullet^T w) \\
&= \int \underbrace{\underbrace{\mathcal{N}(z \mid y_\bullet, \beta^{-1})}_{\text{switch } z \text{ and } y_\bullet} \mathcal{N}(z \mid x_\bullet^T m_N, x_\bullet^T S_N x_\bullet)}_{\text{Use Gaussian product formula SRG-6}} \quad \mathrm{d}z \\
&= \int \mathcal{N}\left(y_\bullet \mid m_N^T x_\bullet, \sigma_N^2(x_\bullet)\right) \underbrace{\mathcal{N}(z \mid \cdot, \cdot)}_{\text{integrate this out}} \quad \mathrm{d}z \\
&= \mathcal{N}\left(y_\bullet \mid m_N^T x_\bullet, \sigma_N^2(x_\bullet)\right)
\end{aligned}
$$

with

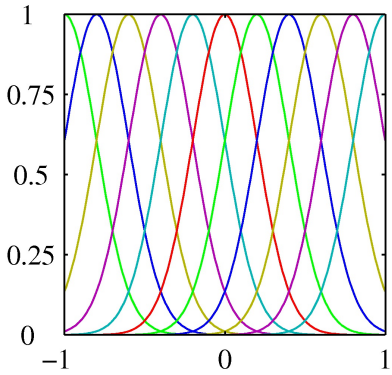
$$
\sigma_N^2(x_\bullet) = \beta^{-1} + x_\bullet^T S_N x_\bullet \tag{B-3.59}
$$

- So, the uncertainty $\sigma_N^2(x_\bullet)$ about the output $y_\bullet$ contains both uncertainty about the process (β^{-1}) and about the model parameters ($x_\bullet^T S_N x_\bullet$).

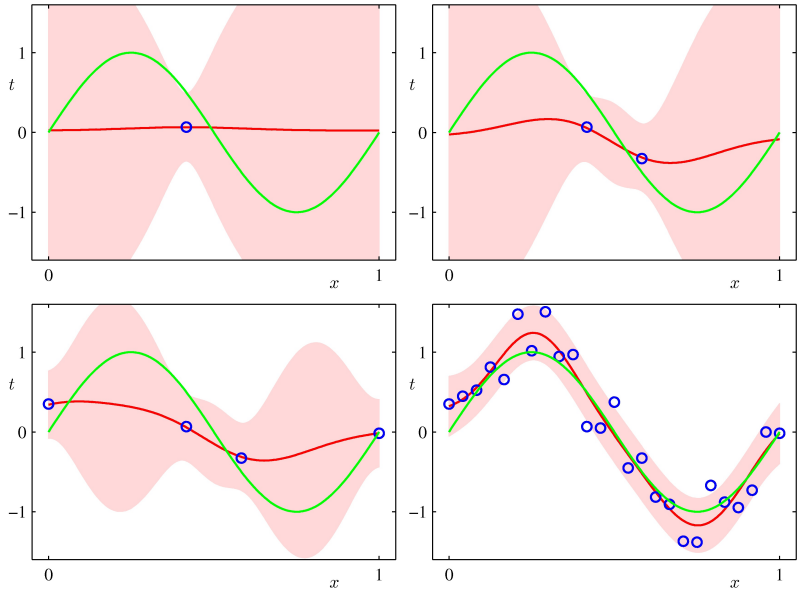
Example Predictive Distribution

- As an example, let's do Bayesian Linear Regression for a synthetic sinusoidal data set and a model with 9 Gaussian basis functions

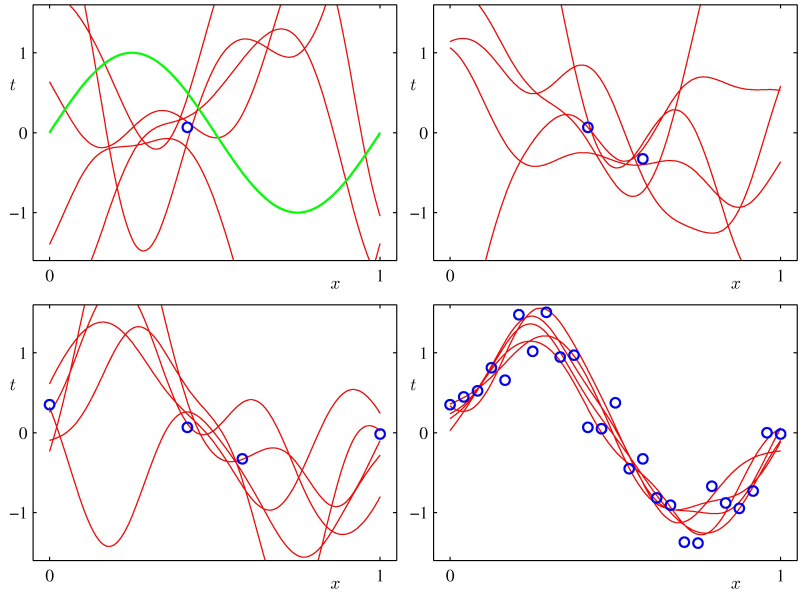
$$
\begin{aligned}
y_n &= \sum_{m=1}^{9} w_m \phi_m(x_n) + \epsilon_n \\
\phi_m(x_n) &= \exp\left(\frac{x_n - \mu_m}{\sigma^2}\right) \\
\epsilon_n &\sim \mathcal{N}(0, \beta^{-1})
\end{aligned}
$$



- The predictive distributions for y are shown in the following plots (Bishop, Fig.3.8)



- And some plots of draws of posteriors for the functions $w^T\phi(x)$ (Bishop, Fig.3.9)



Maximum Likelihood Estimation for Linear Regression Model

- Recall the posterior mean for the weight vector

$$m_N = \left(\frac{\alpha}{\beta}\mathbf{I} + \mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T y$$

where α is the prior precision for the weights.

- The Maximum Likelihood solution for w is obtained by letting $\alpha \to 0$, which leads to

$$\hat{w}_{\text{ML}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T y$$

- The matrix $\mathbf{X}^\dagger \equiv (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ is also known as the **Moore-Penrose pseudo-inverse** (which is sort-of-an-inverse for non-square matrices).

- Note that if we have fewer training samples than input dimensions, i.e., if $N < M$, then $\mathbf{X}^T\mathbf{X}$ will not be invertible and maximum likelihood blows up. The Bayesian solution does not suffer from this problem.

Least-Squares Regression

- (You may say that) we don't need to work with probabilistic models. E.g., there's also the deterministic **least-squares** solution: minimize sum of squared errors,

$$\hat{w}_{\text{LS}} = \arg\min_w \sum_n \left(y_n - w^T x_n\right)^2 = \arg\min_w (y - \mathbf{X}w)^T(y - \mathbf{X}w)$$

- Setting the gradient $\frac{\partial (y-\mathbf{X}w)^T(y-\mathbf{X}w)}{\partial w} = -2\mathbf{X}^T(y - \mathbf{X}w)$ to zero yields the so-called **normal equations** $\mathbf{X}^T\mathbf{X}\hat{w}_{\text{LS}} = \mathbf{X}^T y$ and consequently

$$\hat{w}_{\text{LS}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T y$$

which is the same answer as we got for the maximum likelihood weights w_{ML}.

- $\Rightarrow$ Least-squares regression ($\hat{w}_{\text{LS}}$) corresponds to the (probabilistic) maximum likelihood solution ($\hat{w}_{\text{ML}}$) if the probabilistic model includes the following assumptions:
  1. The observations are independently and identically distributed (**IID**) (this determines how errors are combined), and
  2. The noise signal $\epsilon_n \sim \mathcal{N}(0,\, \beta^{-1})$ is **Gaussian** distributed (determines the error metric)

- If you use the Least-Squares method, you cannot see (nor modify) these assumptions. The probabilistic method forces you to state all assumptions explicitly!

Not Identically Distributed Data

- Let's do an example regarding changing our assumptions. What if we assume that the variance of the measurement error varies with the sampling index, $\epsilon_n \sim \mathcal{N}(0, \beta_n^{-1})$?

- The likelihood is now (using $\Lambda \triangleq \text{diag}(\beta_n)$)

$$p(y|\, \mathbf{X}, w, \Lambda) = \mathcal{N}(y|\, \mathbf{X}w, \Lambda^{-1})\,.$$

- Combining this likelihood with the prior $p(w) = \mathcal{N}(w|\, 0, \alpha^{-1}\mathbf{I})$ leads to a posterior

$$\begin{aligned} p(w|D) &\propto p(D|w)\cdot p(w) \\ &= \mathcal{N}(y|\, \mathbf{X}w, \Lambda^{-1}\mathbf{I}) \cdot \mathcal{N}(w|\, 0, \alpha^{-1}\mathbf{I}) \\ &\propto \exp\left\{\frac{1}{2}(y - \mathbf{X}w)^T\Lambda(y - \mathbf{X}w) + \frac{\alpha}{2}w^T w\right\} \\ &\propto \mathcal{N}(w|\, m_N, S_N) \end{aligned}$$

with

$$\begin{aligned} m_N &= S_N\mathbf{X}^T\Lambda y \\ S_N &= \left(\alpha\mathbf{I} + \mathbf{X}^T\Lambda\mathbf{X}\right)^{-1} \end{aligned}$$

- And maximum likelihood solution

$$w_{\text{ML}} = m_N|_{\alpha\to 0} = \left(\mathbf{X}^T\Lambda\mathbf{X}\right)^{-1}\mathbf{X}^T\Lambda y$$

- This maximum likelihood solution is also called the **Weighted Least Squares** (WLS) solution. (Note that we just stumbled upon it, the crucial aspect is appropriate model specification!)

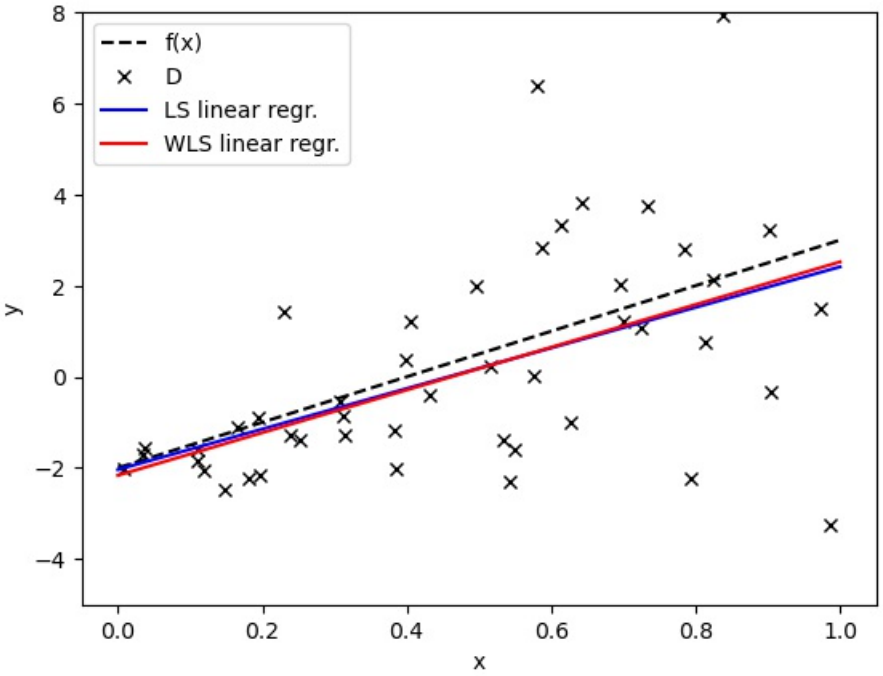
- Note also that the dimension of Λ grows with the number of data points. In general, models for which the number of parameters grow as the number of observations increase are called **non-parametric models**.

Code Example: Least Squares vs Weighted Least Squares

- We'll compare the Least Squares and Weighted Least Squares solutions for a simple linear regression model with input-dependent noise:

$$
\begin{aligned}
x &\sim \text{Unif}[0,1] \\
y|x &\sim \mathcal{N}(f(x), v(x)) \\
f(x) &= 5x - 2 \\
v(x) &= 10e^{2x^2} - 9.5 \\
\mathcal{D} &= \{(x_1, y_1), \ldots, (x_N, y_N)\}
\end{aligned}
$$

In [1]:
```
using PyPlot, LinearAlgebra

# Model specification: y|x ~ 𝒩(f(x), v(x))
f(x) = 5*x .- 2
v(x) = 10*exp.(2*x.^2) .- 9.5 # input dependent noise variance
x_test = [0.0, 1.0]
plot(x_test, f(x_test), "k--") # plot f(x)

# Generate N samples (x,y), where x ~ Unif[0,1]
N = 50
x = rand(N)
y = f(x) + sqrt.(v(x)) .* randn(N)
plot(x, y, "kx"); xlabel("x"); ylabel("y") # Plot samples

# Add constant to input so we can estimate both the offset and the slope
_x = [x ones(N)]
_x_test = hcat(x_test, ones(2))

# LS regression
w_ls = pinv(_x) * y
plot(x_test, _x_test*w_ls, "b-") # plot LS solution

# Weighted LS regression
W = Diagonal(1 ./ v(x)) # weight matrix
w_wls = inv(_x'*W*_x) * _x' * W * y
plot(x_test, _x_test*w_wls, "r-") # plot WLS solution
ylim([-5,8]); legend(["f(x)", "D", "LS linear regr.", "WLS linear regr."],loc=2);
```



OPTIONAL SLIDES

Generative Classification

Preliminaries

- Goal
  - Introduction to linear generative classification with a Gaussian-categorical generative model
- Materials
  - Mandatory

- These lecture notes
- Optional
  - Bishop pp. 196-202 (section 4.2 focusses on binary classification, whereas in these lecture notes we describe generative classification for multiple classes).

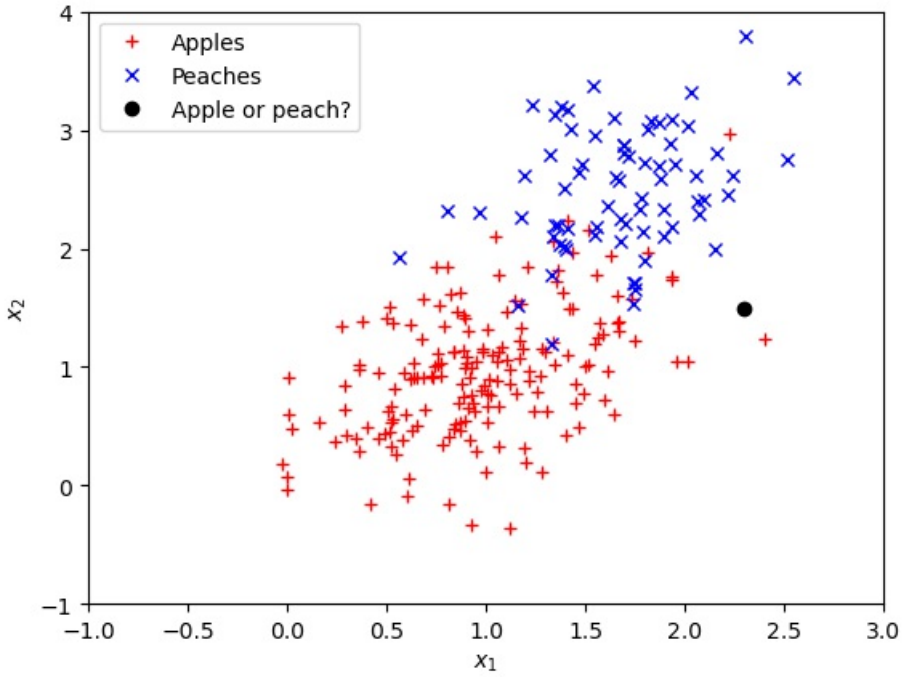
Challenge: an apple or a peach?

- **Problem**: You're given numerical values for the skin features roughness and color for 200 pieces of fruit, where for each piece of fruit you also know if it is an apple or a peach. Now you receive the roughness and color values for a new piece of fruit but you don't get its class label (apple or peach). What is the probability that the new piece is an apple?

- **Solution**: To be solved later in this lesson.

- Let's first generate a data set (see next slide).

In [1]:
```
using Distributions, PyPlot
N = 250; p_apple = 0.7; Σ = [0.2 0.1; 0.1 0.3]
p_given_apple = MvNormal([1.0, 1.0], Σ) # p(X|y=apple)
p_given_peach = MvNormal([1.7, 2.5], Σ) # p(X|y=peach)
X = Matrix{Float64}(undef,2,N); y = Vector{Bool}(undef,N) # true corresponds to apple
for n=1:N
    y[n] = (rand() < p_apple) # Apple or peach?
    X[:,n] = y[n] ? rand(p_given_apple) : rand(p_given_peach) # Sample features
end
X_apples = X[:,findall(y)]'; X_peaches = X[:,findall(.!y)]' # Sort features on class
x_test = [2.3; 1.5] # Features of 'new' data point

function plot_fruit_dataset()
    # Plot the data set and x_test
    plot(X_apples[:,1], X_apples[:,2], "r+")    # apples
    plot(X_peaches[:,1], X_peaches[:,2], "bx") # peaches
    plot(x_test[1], x_test[2], "ko")           # 'new' unlabelled data point
    legend(["Apples"; "Peaches"; "Apple or peach?"], loc=2)
    xlabel(L"x_1"); ylabel(L"x_2"); xlim([-1,3]); ylim([-1,4])
end
plot_fruit_dataset();
```



Generative Classification Problem Statement

- Given is a data set $D = \{(x_1, y_1), \ldots, (x_N, y_N)\}$
  - inputs $x_n \in \mathbb{R}^M$ are called **features**.
  - outputs $y_n \in \mathcal{C}_k$, with $k = 1, \ldots, K$; The **discrete** targets $\mathcal{C}_k$ are called **classes**.

- We will again use the 1-of-K notation for the discrete classes. Define the binary **class selection variable**

$$y_{nk} = \begin{cases} 1 & \text{if } y_n \in \mathcal{C}_k \\ 0 & \text{otherwise} \end{cases}$$

- (Hence, the notations $y_{nk} = 1$ and $y_n \in \mathcal{C}_k$ mean the same thing.)

- The plan for generative classification: build a model for the joint pdf $p(x, y) = p(x|y)p(y)$ and use Bayes to infer the posterior class probabilities

$$p(y|x) = \frac{p(x|y)p(y)}{\sum_{y'} p(x|y')p(y')} \propto p(x|y)\, p(y)$$

1 - Model specification

Likelihood

- Assume Gaussian **class-conditional distributions** with **equal covariance matrix** across the classes,

$$p(x_n|\mathcal{C}_k) = \mathcal{N}(x_n|\mu_k, \Sigma)$$

with notational shorthand: $\mathcal{C}_k \triangleq (y_n \in \mathcal{C}_k)$.

Prior

- We use a categorical distribution for the class labels y_{nk}:

$$p(\mathcal{C}_k) = \pi_k$$

- Hence, using the one-hot coding formulation for y_{nk}, the generative model $p(x_n, y_n)$ can be written as

$$\begin{aligned} p(x_n, y_n) &= \prod_{k=1}^{K} p(x_n, y_{nk} = 1)^{y_{nk}} \\ &= \prod_{k=1}^{K} (\pi_k \cdot \mathcal{N}(x_n|\mu_k, \Sigma))^{y_{nk}} \end{aligned}$$

- We will refer to this model as the **Gaussian-Categorical Model** (GCM).
  - N.B. In the literature, this model (with possibly unequal Σ_k across classes) is often called the Gaussian Discriminant Analysis model and the special case with equal covariance matrices $\Sigma_k = \Sigma$ is also called Linear Discriminant Analysis. We think these names are a bit unfortunate as it may lead to confusion with the discriminative method for classification.

- As usual, once the model has been specified, the rest (inference for parameters and model prediction) through straight probability theory.

Computing the log-likelihood

- The log-likelihood given the full data set $D = \{(x_n, y_n), n = 1, 2, \ldots, N\}$ is then

$$\begin{aligned} \log p(D|\theta) &\overset{\text{IID}}{=} \sum_n \log \prod_k p(x_n, y_{nk} = 1 \mid \theta)^{y_{nk}} \\ &= \sum_{n,k} y_{nk} \log p(x_n, y_{nk} = 1 \mid \theta) \\ &= \sum_{n,k} y_{nk} \log p(x_n|y_{nk} = 1) + \sum_{n,k} y_{nk} \log p(y_{nk} = 1) \\ &= \sum_{n,k} y_{nk} \log \mathcal{N}(x_n|\mu_k, \Sigma) + \sum_{n,k} y_{nk} \log \pi_k \\ &= \sum_{n,k} y_{nk} \underbrace{\log \mathcal{N}(x_n|\mu_k, \Sigma)}_{\text{see Gaussian est.}} + \underbrace{\sum_k m_k \log \pi_k}_{\text{see multinomial est.}} \end{aligned}$$

where we used $m_k \triangleq \sum_n y_{nk}$.

2 - Parameter Inference for Classification

- We'll do Maximum Likelihood estimation for $\theta = \{\pi_k, \mu_k, \Sigma\}$ from data D

- Recall (from the previous slide) the log-likelihood (LLH)

$$\log p(D|\theta) = \sum_{n,k} y_{nk} \underbrace{\log \mathcal{N}(x_n|\mu_k, \Sigma)}_{\text{Gaussian}} + \underbrace{\sum_k m_k \log \pi_k}_{\text{multinomial}}$$

- Maximization of the LLH for the GDA model breaks down into
  - **Gaussian density estimation** for parameters μ_k, Σ, since the first term contains exactly the log-likelihood for MVG density estimation. We've already done this, see the Gaussian distribution lesson.
  - **Multinomial density estimation** for class priors π_k, since the second term holds exactly the log-likelihood for multinomial density estimation, see

the Multinomial distribution lesson.

- The ML for multinomial class prior (we've done this before!)

$$\hat{\pi}_k = \frac{m_k}{N}$$

- Now group the data into separate classes and do MVG ML estimation for class-conditional parameters (we've done this before as well):

$$\begin{aligned}
\mu_k &= \frac{\sum_n y_{nk} x_n}{\sum_n y_{nk}} = \frac{1}{m_k} \sum_n y_{nk} x_n \\
\hat{\Sigma} &= \frac{1}{N} \sum_{n,k} y_{nk} (x_n - \mu_k)(x_n - \mu_k)^T \\
&= \sum_k \pi_k \cdot \underbrace{\left(\frac{1}{m_k} \sum_n y_{nk} (x_n - \mu_k)(x_n - \mu_k)^T \right)}_{\text{class-cond. variance}} \\
&= \sum_k \hat{\pi}_k \cdot \hat{\Sigma}_k
\end{aligned}$$

where $\hat{\pi}_k$, $\hat{\mu}_k$ and $\hat{\Sigma}_k$ are the sample proportion, sample mean and sample variance for the kth class, respectively.

- Note that the binary class selection variable y_{nk} groups data from the same class.

3 - Application: Class prediction for new Data

- Let's apply the trained model to predict the class for given a 'new' input $x_\bullet$:

$$\begin{aligned}
p(\mathcal{C}_k | x_\bullet, D) &= \int p(\mathcal{C}_k | x_\bullet, \theta) \underbrace{p(\theta | D)}_{\text{ML: } \delta(\theta - \hat{\theta})} \, \mathrm{d}\theta \\
&= p(\mathcal{C}_k | x_\bullet, \hat{\theta}) \\
&\propto p(\mathcal{C}_k)\, p(x_\bullet | \mathcal{C}_k) \\
&= \hat{\pi}_k \cdot \mathcal{N}(x_\bullet | \hat{\mu}_k, \hat{\Sigma}) \\
&\propto \hat{\pi}_k \exp\left\{ -\frac{1}{2} (x_\bullet - \hat{\mu}_k)^T \hat{\Sigma}^{-1} (x_\bullet - \hat{\mu}_k) \right\} \\
&= \exp\Big\{ \underbrace{-\frac{1}{2} x_\bullet^T \hat{\Sigma}^{-1} x_\bullet}_{\text{not a function of } k} + \mu_k^T \hat{\Sigma}^{-1} x_\bullet - \frac{1}{2} \mu_k^T \hat{\Sigma}^{-1} \mu_k + \log \pi_k \Big\} \\
&\propto \frac{1}{Z} \exp\{ \beta_k^T x_\bullet + \gamma_k \} \\
&\triangleq \sigma\left(\beta_k^T x_\bullet + \gamma_k \right)
\end{aligned}$$

where $\sigma(a_k) \triangleq \frac{\exp(a_k)}{\sum_{k'} \exp(a_{k'})}$ is called a **softmax** (a.k.a. **normalized exponential**) function, and

$$\begin{aligned}
\beta_k &= \hat{\Sigma}^{-1} \mu_k \\
\gamma_k &= -\frac{1}{2} \mu_k^T \hat{\Sigma}^{-1} \mu_k + \log \hat{\pi}_k \\
Z &= \sum_{k'} \exp\{ \beta_{k'}^T x_\bullet + \gamma_{k'} \}. \quad \text{(normalization constant)}
\end{aligned}$$

- The softmax function is a smooth approximation to the max-function. Note that we did not a priori specify a softmax posterior, but rather it followed from applying Bayes rule to the prior and likelihood assumptions.

- Note the following properties of the softmax function $\sigma(a_k)$:
  - $\sigma(a_k)$ is monotonicaly ascending function and hence it preserves the order of a_k. That is, if $a_j > a_k$ then $\sigma(a_j) > \sigma(a_k)$.
  - $\sigma(a_k)$ is always a proper probability distribution, since $\sigma(a_k) > 0$ and $\sum_k \sigma(a_k) = 1$.

Discrimination Boundaries

- The class log-posterior $\log p(\mathcal{C}_k | x) \propto \beta_k^T x + \gamma_k$ is a linear function of the input features.

- Thus, the contours of equal probability (**discriminant functions**) are lines (hyperplanes) in the feature space

$$\log \frac{p(\mathcal{C}_k | x, \theta)}{p(\mathcal{C}_j | x, \theta)} = \beta_{kj}^T x + \gamma_{kj} = 0$$

where we defined $\beta_{kj} \triangleq \beta_k - \beta_j$ and similarly for γ_{kj}.

- How to classify a new input $x_\bullet$? The Bayesian answer is a posterior distribution $p(\mathcal{C}_k | x_\bullet)$. If you must choose, then the class with maximum posterior class probability

$$\begin{aligned}
k^* &= \arg\max_k p(\mathcal{C}_k | x_\bullet) \\
&= \arg\max_k \left(\beta_k^T x_\bullet + \gamma_k \right)
\end{aligned}$$

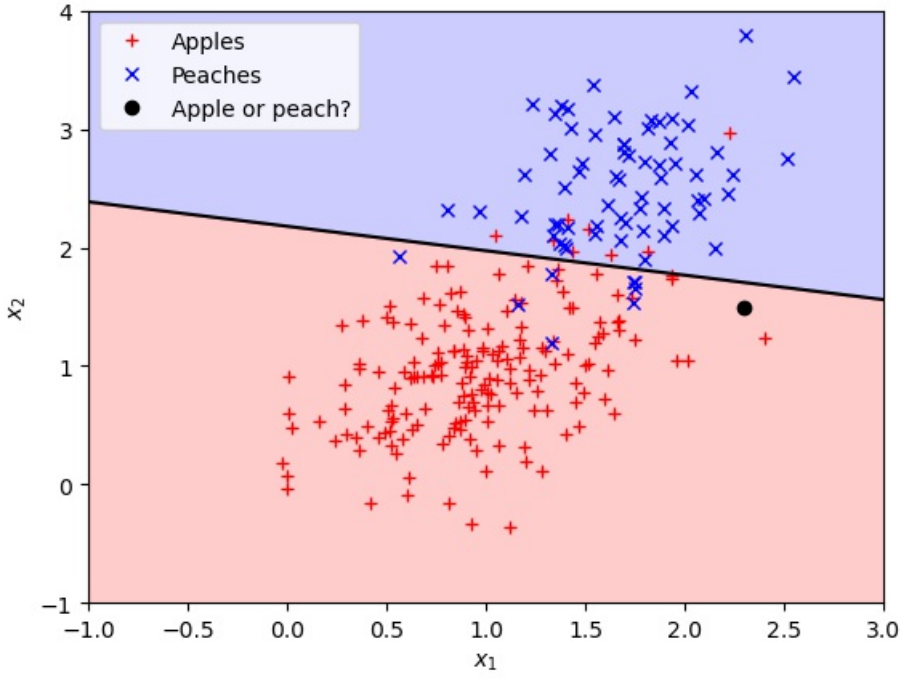
is an appealing decision.

Code Example: Working out the "apple or peach" example problem

We'll apply the above results to solve the "apple or peach" example problem.

In [2]:
```
# Make sure you run the data-generating code cell first

# Multinomial (in this case binomial) density estimation
p_apple_est = sum(y.==true) / length(y)
π_hat = [p_apple_est; 1-p_apple_est]

# Estimate class-conditional multivariate Gaussian densities
d1 = fit_mle(FullNormal, X_apples')  # MLE density estimation d1 = N(μ₁, Σ₁)
d2 = fit_mle(FullNormal, X_peaches') # MLE density estimation d2 = N(μ₂, Σ₂)
Σ = π_hat[1]*cov(d1) + π_hat[2]*cov(d2) # Combine Σ₁ and Σ₂ into Σ
conditionals = [MvNormal(mean(d1), Σ); MvNormal(mean(d2), Σ)] # p(x|C)

# Calculate posterior class probability of x· (prediction)
function predict_class(k, X) # calculate p(Ck|X)
    norm = π_hat[1]*pdf(conditionals[1],X) + π_hat[2]*pdf(conditionals[2],X)
    return π_hat[k]*pdf(conditionals[k], X) ./ norm
end
println("p(apple|x=x·) = $(predict_class(1,x_test))")

# Discrimination boundary of the posterior (p(apple|x;D) = p(peach|x;D) = 0.5)
β(k) = inv(Σ)*mean(conditionals[k])
γ(k) = -0.5 * mean(conditionals[k])' * inv(Σ) * mean(conditionals[k]) + log(π_hat[k])
function discriminant_x2(x1)
    # Solve discriminant equation for x2
    β12 = β(1) .- β(2)
    γ12 = (γ(1) .- γ(2))[1,1]
    return -1*(β12[1]*x1 .+ γ12) ./ β12[2]
end

plot_fruit_dataset() # Plot dataset
x1 = range(-1,length=10,stop=3)
plot(x1, discriminant_x2(x1), "k-") # Plot discrimination boundary
fill_between(x1, -1, discriminant_x2(x1), color="r", alpha=0.2)
fill_between(x1, discriminant_x2(x1), 4, color="b", alpha=0.2);
```

```
p(apple|x=x·) = 0.7429562441220589
```



Recap Generative Classification

- Gaussian-Categorical Model specification:

$$p(x, \mathcal{C}_k|\theta) = \pi_k \cdot \mathcal{N}(x|\mu_k, \Sigma)$$

- If the class-conditional distributions are Gaussian with equal covariance matrices across classes ($\Sigma_k = \Sigma$), then the discriminant functions are hyperplanes in feature space.

- ML estimation for $\{\pi_k, \mu_k, \Sigma\}$ in the GCM model breaks down to simple density estimation for Gaussian and multinomial/categorical distributions.

Discriminative Classification

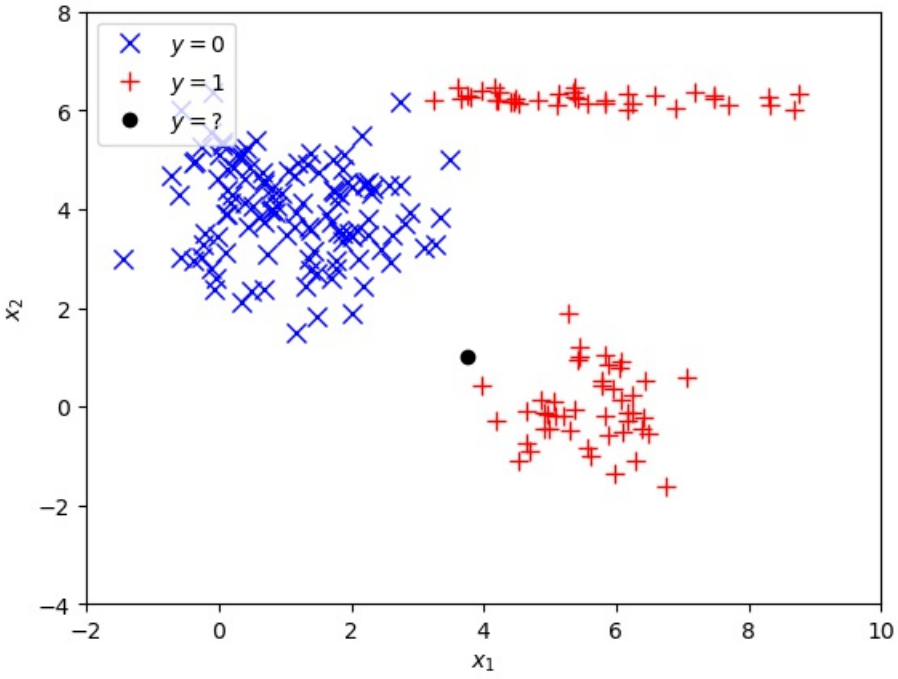
Preliminaries

- Goal
  - Introduction to discriminative classification models
- Materials
  - Mandatory
    - These lecture notes
  - Optional
    - Bishop pp. 213 - 217 (Laplace approximation)
    - Bishop pp. 217 - 220 (Bayesian logistic regression)
    - T. Minka (2005), Discriminative models, not discriminative training

Challenge: difficult class-conditional data distributions

Our task will be the same as in the preceding class on (generative) classification. But this time, the class-conditional data distributions look very non-Gaussian, yet the linear discriminative boundary looks easy enough:

In [1]:
```
using Random; Random.seed!(1234);
# Generate dataset {(x1,y1),...,(xN,yN)}
# x is a 2-d feature vector [x_1;x_2]
# y ∈ {false,true} is a binary class label
# p(x|y) is multi-modal (mixture of uniform and Gaussian distributions)
using PyPlot
include("./scripts/lesson8_helpers.jl")
N = 200
X, y = genDataset(N) # Generate data set, collect in matrix X and vector y
X_c1 = X[:,findall(.!y)]'; X_c2 = X[:,findall(y)]' # Split X based on class label
X_test = [3.75; 1.0] # Features of 'new' data point
function plotDataSet()
    plot(X_c1[:,1], X_c1[:,2], "bx", markersize=8)
    plot(X_c2[:,1], X_c2[:,2], "r+", markersize=8, fillstyle="none")
    plot(X_test[1], X_test[2], "ko")
    xlabel(L"x_1"); ylabel(L"x_2");
    legend([L"y=0", L"y=1",L"y=?"], loc=2)
    xlim([-2;10]); ylim([-4, 8])
end
plotDataSet();
```



Main Idea of Discriminative Classification

- Again, a data set is given by $D = \{(x_1, y_1), \dots, (x_N, y_N)\}$ with $x_n \in \mathbb{R}^M$ and $y_n \in \mathcal{C}_k$, with $k = 1, \dots, K$.
- Sometimes, the precise assumptions of the (Gaussian-Categorical) generative model

  $$p(x_n, y_n \in \mathcal{C}_k | \theta) = \pi_k \cdot \mathcal{N}(x_n | \mu_k, \Sigma)$$

  clearly do not match the data distribution.
- Here's an **IDEA**! Let's model the posterior

  $$p(y_n \in \mathcal{C}_k | x_n)$$

  directly, without any assumptions on the class densities.

Model Specification for Bayesian Logistic Regression

- We will work this idea out for a 2-class problem. Assume a data set is given by $D = \{(x_1, y_1), \dots, (x_N, y_N)\}$ with $x_n \in \mathbb{R}^M$ and $y_n \in \{0, 1\}$.
- What model should we use for the posterior distribution $p(y_n \in \mathcal{C}_k | x_n)$?

Likelihood

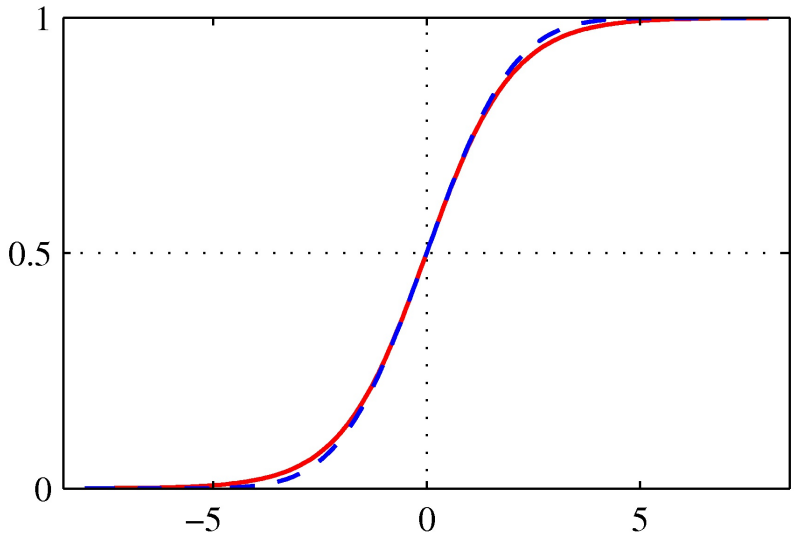
- In Logistic Regression, we take inspiration from the generative approach, where the **softmax** function "emerged" as the posterior. Here, we **choose** the 2-class softmax function (which is called the **logistic** function) with linear discrimination bounderies for the posterior class probability:

  $$p(y_n = 1 \,|\, x_n, w) = \sigma(w^T x_n) \,.$$

  where

  $$\sigma(a) = \frac{1}{1 + e^{-a}}$$

  is the *logistic* function.
- Clearly, it follows from this assumption that $p(y_n = 0 \,|\, x_n, w) = 1 - \sigma(w^T x_n)$.



- (Bishop fig.4.9). The logistic function $\sigma(a) = 1/(1+e^{-a})$ (red), together with the scaled probit function $\Phi(\lambda a)$, for $\lambda^2 = \pi/8$ (in blue). We will use this approximation later in the Laplace approximation.

- Adding the other class ($y_n = 0$) leads to the following posterior class distribution:

$$\begin{aligned} p(y_n \mid x_n, w) &= \text{Bernoulli}\left(y_n \mid \sigma(w^T x_n)\right) \\ &= \sigma(w^T x_n)^{y_n}\left(1-\sigma(w^T x_n)\right)^{(1-y_n)} \\ &= \sigma\left((2y_n-1)w^T x_n\right) \end{aligned} \tag{B-4.89}$$

  - Note that for the 3rd equality, we have made use of the fact that $\sigma(-a) = 1-\sigma(a)$.
  - Each of these three models in B-4.89 are **equivalent**. We mention all three notational options since they all appear in the literature.

- For the data set $D = \{(x_1, y_1), \ldots, (x_N, y_N)\}$, the **likelihood function** for the parameters w is then given by

$$p(D|w) = \prod_{n=1}^{N} \sigma\left((2y_n-1)w^T x_n\right)$$

- This choice for the class posterior is called **logistic regression**, in analogy to **linear regression**:

$$\begin{aligned} p(y_n|x_n, w) &= \mathcal{N}(y_n|w^T x_n, \beta^{-1}) && \text{for linear regression} \\ p(y_n|x_n, w) &= \sigma\left((2y_n-1)w^T x_n\right) && \text{for logistic regression} \end{aligned}$$

- In the discriminative approach, the parameters w are **not** structured into $\{\mu, \Sigma, \pi\}$. In principle they are "free" parameters for which we can choose any value that seems appropriate. This provides discriminative approach with more flexibility than the generative approach.

Prior

- In *Bayesian* logistic regression, we often add a **Gaussian prior on the weights**:

$$p(w) = \mathcal{N}(w|\, m_0, S_0) \tag{B-4.140}$$

Some Notes on the Model

- Note that for generative classification, for the sake of simplicity, we used maximum likelihood estimation for the model parameters. In this lesson on discriminative classification, we specify both a prior and likelihood function for the parameters w, which allows us to compute a Bayesian posterior for the weights. In principle, we could have used Bayesian parameter estimation for the generative classification model as well (but the math is not suited for a introductory lesson).

- In the optional paper by T. Minka (2005), you can read how the model assumptions for discriminative classification can be re-interpreted as a special generative model (this paper not for exam).

- As an exercise, please check that for logistic regression with $p(y_n = 1 \mid x_n, w) = \sigma(w^T x_n)$, the **discrimination boundary**, which can be computed by

$$\frac{p(y_n \in \mathcal{C}_1|x_n)}{p(y_n \in \mathcal{C}_0|x_n)} \overset{!}{=} 1$$

  is a straight line, see Exercises.

Inference

- After model specification, the rest follows by application of probability theory.
- The posterior for the weights follows by Bayes rule

$$\begin{aligned}\underbrace{p(w|\,D)}_{\text{posterior}} &\propto p(w)p(D|w)\\ &= \underbrace{\mathcal{N}(w|\,m_0,\,S_0)}_{\text{prior}} \cdot \underbrace{\prod_{n=1}^{N} \sigma\left((2y_n-1)w^T x_n\right)}_{\text{likelihood}}\end{aligned} \tag{B-4.142}$$

- In principle, Bayesian inference is done now!
- Unfortunately, the posterior $p(w|\,D)$ is not Gaussian and the evidence $p(D)$ is also not analytically computable. (We will deal with this later).

Predictive distribution

- For a new data point $x_\bullet$, the predictive distribution for $y_\bullet$ is given by

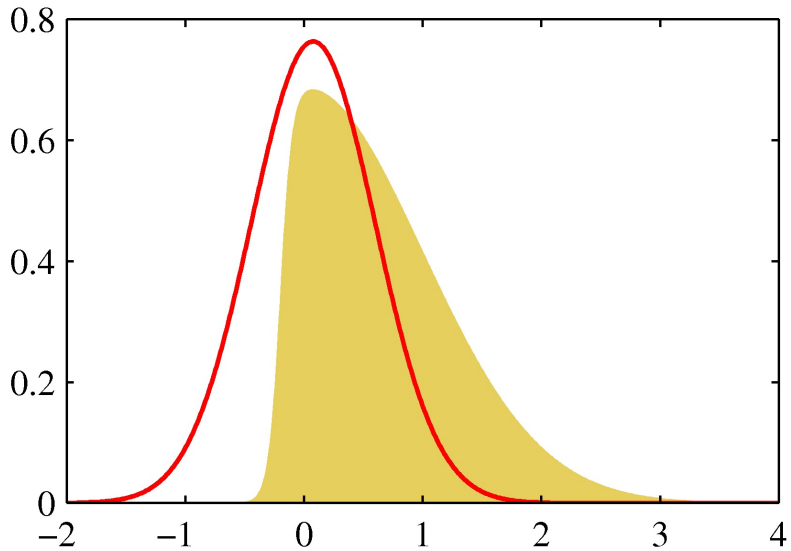
$$\begin{aligned}p(y_\bullet = 1\,|\,x_\bullet,\,D) &= \int p(y_\bullet = 1\,|\,x_\bullet, w)\; p(w|\,D)\,\mathrm{d}w\\ &= \int \sigma(w^T x_\bullet)\, p(w\,|\,D)\,\mathrm{d}w\end{aligned} \tag{B-4.145}$$

- After substitution of $p(w|D)$ from B-4.142, we have closed-form expressions for both the posterior $p(w|D)$, evidence $p(D)$ and the predictive distribution $p(y_\bullet = 1\,|\,x_\bullet, D)$. Unfortunately, these expressions are not analytically computable.
- Many methods have been developed to approximate the integrals for the predictive distribution and evidence. Here, we present the **Laplace approximation**, which is one of the simplest methods with broad applicability to Bayesian calculations.

The Laplace Approximation

- The central idea of the Laplace approximation is to approximate a (possibly unnormalized) distribution $f(z)$ by a Gaussian distribution $q(z)$.
- Note that $\log q(z)$ is a second-order polynomial in z, so we will find the Gaussian by fitting a parabola to $\log f(z)$.

Example



- (Bishop fig.4.14a). Laplace approximation (in red) to the distribution $p(z) \propto \exp(-z^2/2)\sigma(20z+4)$, where $\sigma(a) = 1/(1+e^{-a})$. The Laplace approximation is centered on the mode of $p(z)$.

Working out the Laplace Approximation

estimation of mean

- The mean (z_0) of $q(z)$ is placed on the mode of $\log f(z)$, i.e.,

$$z_0 = \arg\max_z \log f(z) \tag{B-4.126}$$

estimation of precision matrix

- Note that since $\nabla \log f(z) = \frac{1}{f(z)} \nabla f(z)$ and the gradient $\nabla f(z)|_{z=z_0}$ vanishes at the mode $z = z_0$, we can (Taylor) expand $\log f(z)$ around $z = z_0$ as

$$\begin{aligned} \log f(z) &\approx \log f(z_0) + \overbrace{(\nabla \log f(z_0))^T (z - z_0)}^{=0 \text{ at } z=z_0} + \dots \\ &\quad + \frac{1}{2}(z - z_0)^T \left(\nabla\nabla \log f(z_0)\right)(z - z_0) \\ &= \log f(z_0) - \frac{1}{2}(z - z_0)^T A (z - z_0) \end{aligned} \tag{B-4.131}$$

  where the Hessian matrix A is defined by

$$A = -\nabla\nabla \log f(z)|_{z=z_0} \tag{B-4.132}$$

Laplace approximation construction

- After taking exponentials in eq. B-4.131, we obtain

$$f(z) \approx f(z_0) \exp\left(-\frac{1}{2}(z - z_0)^T A (z - z_0)\right)$$

- We can now identify $q(z)$ as

$$q(z) = \mathcal{N}\left(z \,|\, z_0, A^{-1}\right) \tag{B-4.134}$$

  with z_0 and A defined by eqs. B-4.126 and B-4.132.

- All we have done up to now is approximate a function $f(z)$ by a Gaussian $q(z)$. This procedure is called the **Laplace Approximation**. Often, the required integrals (for Bayesian marginalization) can be approximately computed if we replace $f(z)$ by $q(z)$.

Bayesian Logistic Regression with the Laplace Approximation

- Let's get back to the challenge of computing the predictive class distribution (B-4.145) for Bayesian logistic regression. We first work out the Gaussian Laplace approximation $q(w)$ to the posterior weight distribution

$$\underbrace{p(w|D)}_{\text{posterior}} \propto \underbrace{\mathcal{N}(w|\, m_0, S_0)}_{\text{prior}} \cdot \underbrace{\prod_{n=1}^{N} \sigma\left((2y_n - 1) w^T x_n\right)}_{\text{likelihood}} \tag{B-4.142}$$

A Gausian Laplace approximation to the weights posterior

- Since we have a differentiable expression for $\log p(w|D)$, it is straightforward to compute the gradient and Hessian (for proof, see optional slide):

$$\begin{aligned} \nabla_w \log p(w|D) &= S_0^{-1} \cdot (m_0 - w) + \sum_n (2y_n - 1)(1 - \sigma_n) x_n \\ \nabla\nabla_w \log p(w|D) &= -S_0^{-1} - \sum_n \sigma_n (1 - \sigma_n) x_n x_n^T \end{aligned} \tag{B-4.143}$$

  where we used shorthand σ_n for $\sigma\left((2y_n - 1) w^T x_n\right)$.

- We can now use the gradient $\nabla_w \log p(w|D)$ to find the **mode** w_N of $\log p(w|D)$ (eg by some gradient-based optimization procedure) and then use the Hessian $\nabla\nabla_w \log p(w|D)$ to get the variance of $q(w)$, leading to a **Gaussian approximate weights posterior**:

$$q(w) = \mathcal{N}(w \,|\, w_N, S_N) \tag{B-4.144}$$

  with

$$S_N^{-1} = S_0^{-1} + \sum_n \sigma_n (1 - \sigma_n) x_n x_n^T \tag{B-4.143}$$

Using the Laplace-approximated parameter posterior to evaluate the predictive distribution

- In the analytically unsolveable expressions for evidence and the predictive distribution (estimating the class of a new observation), we proceed with using the Laplace approximation to the weights posterior. For a new observation $x_\bullet$, the class probability is now

$$
\begin{aligned}
p(y_\bullet = 1 \mid x_\bullet, D) &= \int p(y_\bullet = 1 \mid x_\bullet, w) \cdot p(w \mid D) \, \mathrm{d}w \\
&\approx \int p(y_\bullet = 1 \mid x_\bullet, w) \cdot \underbrace{q(w)}_{\text{Gaussian}} \, \mathrm{d}w \\
&= \int \sigma(w^T x_\bullet) \cdot \mathcal{N}(w \mid w_N, S_N) \, \mathrm{d}w
\end{aligned} \tag{B-4.145}
$$

- This looks better but we need two more clever tricks to evaluate this expression.
  1. First, note that w appears in $\sigma(w^T x_\bullet)$ as an inner product, so through substitution of $a := w^T x_\bullet$, the expression simplifies to an integral over the scalar a (see Bishop for derivation):

$$
p(y_\bullet = 1 \mid x_\bullet, D) \approx \int \sigma(a) \, \mathcal{N}\left(a \mid \mu_a, \sigma_a^2\right) \mathrm{d}a \tag{B-4.151}
$$

$$
\mu_a = w_N^T x_\bullet \tag{B-4.149}
$$

$$
\sigma_a^2 = x_\bullet^T S_N x_\bullet \tag{B-4.150}
$$

  2. Secondly, while the integral of the product of a logistic function with a Gaussian is not analytically solvable, the integral of the product of a Gaussian cumulative distribution function (CDF, also known as the probit function) with a Gaussian *does* have a closed-form solution. Fortunately,

$$
\Phi(\lambda a) \approx \sigma(a)
$$

  with the Gaussian CDF $\Phi(x) = \frac{1}{\sqrt{(2\pi)}} \int_{-\infty}^{x} e^{-t^2/2} \mathrm{d}t$, $\lambda^2 = \pi/8$ and $\sigma(a) = 1/(1+e^{-a})$. Thus, substituting $\Phi(\lambda a)$ with $\lambda^2 = \pi/8$ for $\sigma(a)$ leads to

$$
\begin{aligned}
p(y_\bullet = 1 \mid x_\bullet, D) &= \int \sigma(w^T x_\bullet) \cdot p(w|D) \, \mathrm{d}w \\
&\approx \int \underbrace{\Phi(\lambda a)}_{\text{probit function}} \cdot \underbrace{\mathcal{N}\left(a \mid \mu_a, \sigma_a^2\right)}_{\text{Gaussian}} \, \mathrm{d}a \\
&= \Phi\left(\frac{\mu_a}{\sqrt{(\lambda^{-2} + \sigma_a^2)}} \right)
\end{aligned} \tag{B-4.152}
$$

- We now have an approximate but **closed-form expression for the predictive class distribution for a new observation** with a Bayesian logistic regression model.

- Note that, by Eq.B-4.143, the variance S_N (and consequently σ_a^2) for the weight vector depends on the distribution of the training set. Large uncertainty about the weights (in areas with little training data and uninformative prior variance S_0) increases σ_a^2 and takes the posterior class probability eq. B-4.152 closer to 0.5. Does that make sense?

- Apparently, the Laplace approximation leads to a closed-form solutions for Bayesian logistic regression (although admittedly, the derivation is no walk in the park).

- Exam guide: The derivation of closed-form expression eq. B-4.152 for the predictive class distribution requires clever tricks and is therefore not something that you should be able to reproduce at the exam without assistance. You should understand the Laplace Approximation though and be able to work out simpler examples.

ML Estimation for Discriminative Classification

- Rather than the computationally involved Laplace approximation for Bayesian inference, in practice, discriminative classification is often executed through maximum likelihood estimation.

- With the usual 1-of-K encoding scheme for classes ($y_{nk} = 1$ if $x_n \in \mathcal{C}_k$, otherwise $y_{nk} = 0$), the log-likelihood for a K-dimensional discriminative classifier is

$$
\begin{aligned}
\mathrm{L}(\theta) &= \log \prod_n \prod_k p(\mathcal{C}_k|x_n, \theta)^{y_{nk}} \\
&= \log \prod_n \prod_k \Bigg(\underbrace{\frac{e^{\theta_k^T x_n}}{\sum_j e^{\theta_j^T x_n}}}_{\text{softmax function}} \Bigg)^{y_{nk}} \\
&= \sum_n \sum_k y_{kn} \log \left(\frac{e^{\theta_k^T x_n}}{\sum_j e^{\theta_j^T x_n}} \right)
\end{aligned}
$$

- Computing the gradient $\nabla_{\theta_k} \mathrm{L}(\theta)$ leads to (for proof, see optional slide below)

$$
\nabla_{\theta_k} \mathrm{L}(\theta) = \sum_n \underbrace{\Bigg(\underbrace{y_{nk}}_{\text{target}} - \underbrace{\frac{e^{\theta_k^T x_n}}{\sum_j e^{\theta_j^T x_n}}}_{\text{prediction}} \Bigg)}_{\text{prediction error}} \cdot x_n
$$

- Compare this to the gradient for *linear* regression:

$$
\nabla_\theta \mathrm{L}(\theta) = \sum_n \left(y_n - \theta^T x_n \right) x_n
$$

- In both cases

$$\nabla_\theta \mathrm{L} = \sum_n (\mathrm{target}_n - \mathrm{prediction}_n) \cdot \mathrm{input}_n$$

- The parameter vector θ for logistic regression can be estimated through iterative gradient-based adaptation. E.g. (with iteration index i),

$$\theta^{(i+1)} = \theta^{(i)} + \eta \cdot \nabla_\theta \mathrm{L}(\theta)|_{\theta=\hat{\theta}^{(i)}}$$

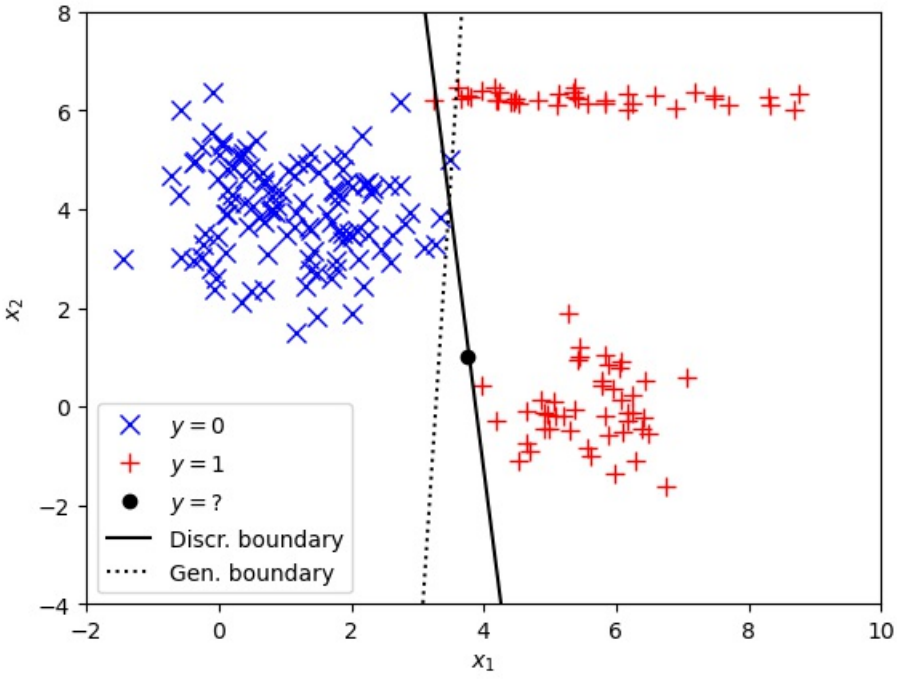
  - Note that, while in the Bayesian approach we get to update θ with **precision-weighted** prediction errors (which is optimal), in the maximum likelihood approach, we weigh the prediction errors with **input** values (which is less precise).

Code Example: ML Estimation for Discriminative Classification

- Let us perform ML estimation of w on the data set from the introduction. To allow an offset in the discrimination boundary, we add a constant 1 to the feature vector x. We only have to specify the (negative) log-likelihood and the gradient w.r.t. w. Then, we use an off-the-shelf optimisation library to minimize the negative log-likelihood.
- We plot the resulting maximum likelihood discrimination boundary. For comparison we also plot the ML discrimination boundary obtained from the code example in the generative Gaussian classifier lesson.

In [2]:
```
using Optim # Optimization library

    y_1 = zeros(length(y))# class 1 indicator vector
    y_1[findall(y)] .= 1
    X_ext = vcat(X, ones(1, length(y))) # Extend X with a row of ones to allow an offset in the discrimination boundary

    # Implement negative log-likelihood function
    function negative_log_likelihood(θ::Vector)
        # Return negative log-likelihood: -L(θ)
        p_1 = 1.0 ./ (1.0 .+ exp.(-X_ext' * θ))    # P(C1|X,θ)
        return -sum(log.( (y_1 .* p_1) + ((1 .- y_1).*(1 .- p_1))) ) # negative log-likelihood
    end

    # Use Optim.jl optimiser to minimize the negative log-likelihood function w.r.t. θ
    results = optimize(negative_log_likelihood, zeros(3), LBFGS())
    θ = results.minimizer

    # Plot the data set and ML discrimination boundary
    plotDataSet()
    p_1(x) = 1.0 ./ (1.0 .+ exp(-([x;1.]' * θ)))
    boundary(x1) = -1 ./ θ[2] * (θ[1]*x1 .+ θ[3])
    plot([-2.;10.], boundary([-2.; 10.]), "k-");
    # # Also fit the generative Gaussian model from lesson 7 and plot the resulting discrimination boundary for compariso
    generative_boundary = buildGenerativeDiscriminationBoundary(X, y)
    plot([-2.;10.], generative_boundary([-2;10]), "k:");
    legend([L"y=0";L"y=1";L"y=?";"Discr. boundary";"Gen. boundary"], loc=3);

    # Given $\hat{\theta}$, we can classify a new input $x_\bullet = [3.75, 1.0]^T$:

    x_test = [3.75;1.0]
    println("P(C1|x•,θ) = $(p_1(x_test))")
```

```
P(C1|x•,θ) = 0.4016963543060162
```



- The generative model gives a bad result because the feature distribution of one class is clearly non-Gaussian: the model does not fit the data well.
- The discriminative approach does not suffer from this problem because it makes no assumptions about the feature distribition $p(x|y)$, it just estimates the conditional class distribution $p(y|x)$ directly.

Recap Classification

	Generative	Discriminative (ML)
1	Like **density estimation**, model joint prob. $p(\mathcal{C}_k)p(x\|\mathcal{C}_k) = \pi_k \mathcal{N}(\mu_k, \Sigma)$	Like (linear) **regression**, model conditional $p(\mathcal{C}_k\|x, \theta)$
2	Leads to **softmax** posterior class probability $p(\mathcal{C}_k\|x, \theta) = e^{\theta_k^T x}/Z$ with **structured** θ	**Choose** also softmax posterior class probability $p(\mathcal{C}_k\|x, \theta) = e^{\theta_k^T x}/Z$ but now with 'free' θ
3	For Gaussian $p(x\|\mathcal{C}_k)$ and multinomial priors, $\theta_k = \begin{bmatrix} -\frac{1}{2}\mu_k^T \sigma^{-1} \mu_k + \log \pi_k \\ \sigma^{-1}\mu_k \end{bmatrix}$ **in one shot**.	Find θ_k through gradient-based adaptation $\nabla_{\theta_k} L(\theta) = \sum_n \left(y_{nk} - \frac{e^{\theta_k^T x_n}}{\sum_{k'} e^{\theta_{k'}^T x_n}}\right) x_n$

OPTIONAL SLIDES

Proof of gradient and Hessian for Laplace Approximation of Posterior

- We will start with the posterior

$$\underbrace{p(w|D)}_{\text{posterior}} \propto \underbrace{\mathcal{N}(w|\, m_0, S_0)}_{\text{prior}} \cdot \underbrace{\prod_{n=1}^{N} \sigma(\underbrace{(2y_n - 1)w^T x_n}_{a_n})}_{\text{likelihood}} \tag{B-4.142}$$

from which it follows that

$$\log p(w|D) \propto -\frac{1}{2}\log|S_0| - \frac{1}{2}(w - m_0)^T S_0^{-1}(w - m_0) + \sum_n \log \sigma(a_n)$$

and the gradient

$$\nabla_w \log p(w|D) \propto \underbrace{S_0^{-1}(m_0 - w)}_{\text{SRM-5b}} + \sum_n \underbrace{\frac{1}{\sigma(a_n)}}_{\frac{\partial \log \sigma(a_n)}{\partial \sigma(a_n)}} \cdot \underbrace{\sigma(a_n) \cdot (1 - \sigma(a_n))}_{\frac{\partial \sigma(a_n)}{\partial a_n}} \cdot \underbrace{(2y_n - 1)x_n}_{\frac{\partial a_n}{\partial w} \text{ (see SRM-5a)}}$$

$$= S_0^{-1}(m_0 - w) + \sum_n (2y_n - 1)(1 - \sigma(a_n))x_n \quad \text{(gradient)}$$

where we used $\sigma'(a) = \sigma(a) \cdot (1 - \sigma(a))$.

- For the Hessian, we continue to differentiate the transpose of the gradient, leading to

$$\begin{aligned}
\nabla\nabla_w \log p(w|D) &= \nabla_w \left(S_0^{-1}(m_0 - w) \right)^T - \sum_n (2y_n - 1) x_n \nabla_w \sigma(a_n)^T \\
&= -S_0^{-1} - \sum_n (2y_n - 1) x_n \cdot \underbrace{\sigma(a_n) \cdot (1 - \sigma(a_n))}_{\frac{\partial \sigma(a_n)^T}{\partial a_n^T}} \cdot \underbrace{(2y_n - 1) x_n^T}_{\frac{\partial a_n^T}{\partial w}} \\
&= -S_0^{-1} - \sum_n \sigma(a_n) \cdot (1 - \sigma(a_n)) \cdot x_n x_n^T \quad \text{(Hessian)}
\end{aligned}$$

since $(2y_n - 1)^2 = 1$ for $y_n \in \{0, 1\}$.

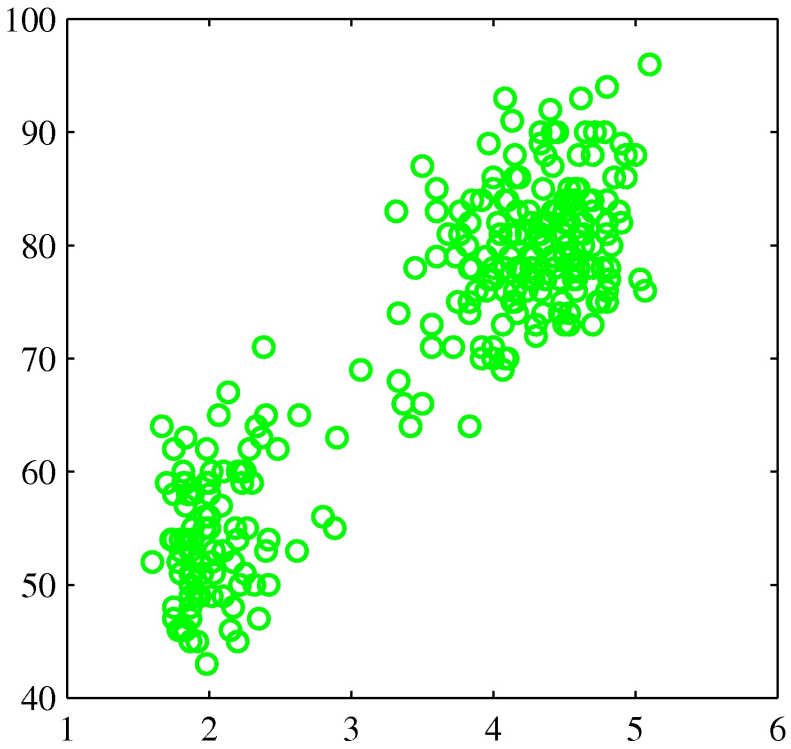
Latent Variable Models and Variational Bayes

Preliminaries

- Goal
  - Introduction to latent variable models and variational inference by Free energy minimization
- Materials
  - Mandatory
    - These lecture notes
    - Ariel Caticha (2010), Entropic Inference
      - tutorial on entropic inference, which is a generalization to Bayes rule and provides a foundation for variational inference.
  - Optional
    - Bishop (2016), pp. 461-486 (sections 10.1, 10.2 and 10.3)
  - references
    - Blei et al. (2017), Variational Inference: A Review for Statisticians
    - Lanczos (1961), The variational principles of mechanics
    - Senoz et al. (2021), Variational Message Passing and Local Constraint Manipulation in Factor Graphs
    - Dauwels (2007), On variational message passing on factor graphs
    - Caticha (2010), Entropic Inference
    - Shore and Johnson (1980), Axiomatic Derivation of the Principle of Maximum Entropy and the Principle of Minimum Cross-Entropy

Illustrative Example : Density Modeling for the Old Faithful Data Set

- You're now asked to build a density model for a data set (Old Faithful, Bishop pg. 681) that clearly is not distributed as a single Gaussian:



Unobserved Classes

Consider again a set of observed data $D = \{x_1, \ldots, x_N\}$

- This time we suspect that there are *unobserved* class labels that would help explain (or predict) the data, e.g.,
  - the observed data are the color of living things; the unobserved classes are animals and plants.
  - observed are wheel sizes; unobserved categories are trucks and personal cars.
  - observed is an audio signal; unobserved classes include speech, music, traffic noise, etc.

- Classification problems with unobserved classes are called **Clustering** problems. The learning algorithm needs to **discover the classes from the observed data**.

The Gaussian Mixture Model

- The spread of the data in the illustrative example looks like it could be modeled by two Gaussians. Let's develop a model for this data set.

- Let $D = \{x_n\}$ be a set of observations. We associate a one-hot coded hidden class label z_n with each observation:

$$z_{nk} = \begin{cases} 1 & \text{if } x_n \in \mathcal{C}_k \,(\text{the k-th class}) \\ 0 & \text{otherwise} \end{cases}$$

- We consider the same model as we did in the generative classification lesson:

$$\begin{aligned} p(x_n | z_{nk} = 1) &= \mathcal{N}(x_n | \mu_k, \Sigma_k) \\ p(z_{nk} = 1) &= \pi_k \end{aligned}$$

  which can be summarized with the selection variables z_{nk} as

$$p(x_n, z_n) = p(x_n | z_n) p(z_n) = \prod_{k=1}^{K} (\underbrace{\pi_k \cdot \mathcal{N}(x_n | \mu_k, \Sigma_k)}_{p(x_n, z_{nk}=1)})^{z_{nk}}$$

- *Again*, this is the same model as we defined for the generative classification model: A Gaussian-Categorical model but now with unobserved classes).
- This model (with **unobserved class labels**) is known as a **Gaussian Mixture Model** (GMM).

The Marginal Distribution for the GMM

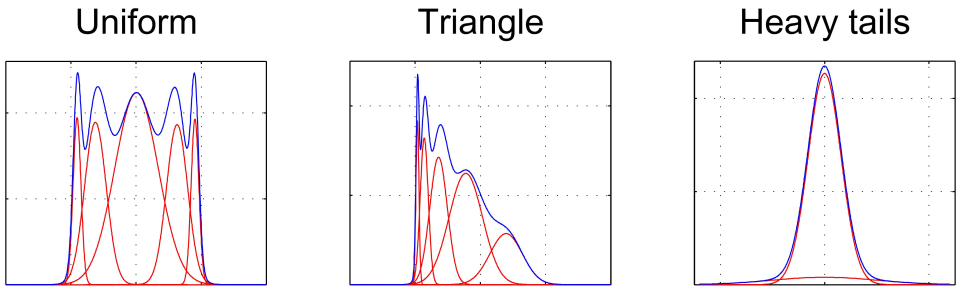
- In the literature, the GMM is often introduced by the marginal distribution for an *observed* data point x_n, given by

$$p(x_n) = \sum_{z_n} p(x_n, z_n)$$
$$= \sum_{k=1}^{K} \pi_k \cdot \mathcal{N}(x_n | \mu_k, \Sigma_k) \quad \text{(B-9.12)}$$

- Full proof as an exercise.
- Note that Eq. B-9.12 is not the generative model. The generative model is the joint distribution $p(x, z)$ over all variables, including the latent variables.
- Eq. B-9.12 reveals the link to the name Gaussian *mixture model*. The priors π_k are also called **mixture coefficients**.

GMM is a very flexible model

- GMMs are very popular models. They have decent computational properties and are **universal approximators of densities** (as long as there are enough Gaussians of course)



- (In the above figure, the Gaussian components are shown in red and the pdf of the mixture models in blue).

Latent Variable Models

- The GMM contains both *observed* variables $\{x_n\}$, (unobserved) *parameters* $\theta = \{\pi_k, \mu_k, \Sigma_k\}$ *and* unobserved (synonym: latent, hidden) variables $\{z_{nk}\}$.
- From a Bayesian viewpoint, both latent variables $\{z_{nk}\}$ and parameters θ are just unobserved variables for which we can set a prior and compute a posterior by Bayes rule.
- Note that z_{nk} has a subscript n, hence its value depends not only on the cluster (k) but also on the n-th observation (in contrast to parameters). These observation-dependent latent variables are generally a useful tool to encode structure in the model. Here (in the GMM), the latent variables $\{z_{nk}\}$ encode (unobserved) class membership.
- By adding model structure through (equations among) latent variables, we can build very complex models. Unfortunately, inference in latent variable models can also be much more complex than for fully observed models.

Inference for GMM is Difficult

- We recall here the log-likelihood for the Gaussian-Categorial Model, see generative classification lesson)

$$\log p(D|\theta) = \sum_{n,k} y_{nk} \underbrace{\log \mathcal{N}(x_n | \mu_k, \Sigma)}_{\text{Gaussian}} + \underbrace{\sum_{n,k} y_{nk} \log \pi_k}_{\text{multinomial}} .$$

- Since the class labels $y_{nk} \in \{0, 1\}$ were given, this expression decomposed into a set of simple updates for the Gaussian and multinomial distributions. For both distributions, we have conjugate priors, so inference is easily solved.
- However, for the Gaussian mixture model (same log-likelihood function with z_{nk} replacing y_{nk}), the class labels $\{z_{nk}\}$ are *unobserved* and need to estimated alongside with the parameters.
- There is no conjugate prior on the latent variables for the GMM likelihood function $p(\underbrace{D}_{\{x_n\}} | \underbrace{\{z_{nk}\}, \{\mu_k, \Sigma_k, \pi_k\}}_{\text{all latent variables}})$.
- In this lesson, we introduce an approximate Bayesian inference method known as **Variational Bayes** (VB) (also known as **Variational Inference**) that can be used for Bayesian inference in models with latent variables. Later in this lesson, we will use VB to do inference in the GMM.
- As a motivation for variational inference, we first discuss the foundation of inference by the **Method of Maximum Relative Entropy**, (Caticha, 2010).

Inference When Information is in the Form of Constraints

- In the probability theory lesson, we recognized Bayes rule as the fundamental rule for learning from data.

- We will now generalize this notion and consider learning as a process that updates a prior into a posterior distribution whenever new information becomes available.

- In this context, *new information* is not necessarily a new observation, but could (for instance) also relate to adding *constraints* on the posterior distribution.

- For example, consider a model $\prod_n p(x_n, z) = p(z) \prod_n p(x_n|z)$.

- Our prior beliefs about z are represented by $p(z)$. In the following, we will write $q(z)$ to denote a posterior distribution for z.

- We might be interested in obtaining rational posterior beliefs $q(z)$ in consideration of the following additional constraints:
  1. *Data Constraints*: e.g., two observed data points $x_1 = 5$ and $x_2 = 3$, which lead to constraints $q(x_1) = \delta(x_1 - 5)$ and $q(x_2) = \delta(x_2 - 3)$.
  2. *Form Constraints*: e.g., we only consider the Gamma distribution for $q(z) = \text{Gam}(z|\alpha, \beta)$.
  3. *Factorization Constraints*:, e.g., we consider independent marginals for the posterior distribution: $q(z) = \prod_k q(z_k)$.
  4. *Moment Constraints*: e.g., the first moment of the posterior is given by $\int z q(z) \text{d}z = 3$.

- Note that this is not "just" applying Bayes rule to compute a posterior, which can only deal with data constraints as specified by the likelihood function.

- Note also that observations *can* be rephrased as constraints on the posterior, e.g., observation $x_1 = 5$ can be phrased as a constraint $q(x_1) = \delta(x_1 - 5)$.

- $\Rightarrow$ Updating a prior to a posterior on the basis of constraints on the posterior is more general than updating based on observations alone.

The Method of Maximum Relative Entropy

- Caticha (2010) (based on earlier work by Shore and Johnson (1980)) developed the **method of maximum (relative) entropy** for rational updating of priors to posteriors when faced with new information in the form of constraints.

- Consider prior beliefs $p(z)$ about z. New information in the form of constraints is obtained and we are interested in the "best update" to a posterior $q(z)$.

- In order to define what "best update" means, we assume a ranking function $S[q, p]$ that generates a preference score for each candidate posterior q for a given p. The best update from p to q is identified as $q^* = \arg\max_q S[q, p]$.

- Similarly to Cox' method to deriving probability theory, Caticha introduced the following mild criteria based on a rational principle (the **principle of minimal updating**, see Caticha 2010) that the ranking function needs to adhere to:
  1. *Locality*: local information has local effects.
  2. *Coordinate invariance*: the system of coordinates carries no information.
  3. *Independence*: When systems are known to be independent, it should not matter whether they are treated separately or jointly.

- These three criteria **uniquely identify the Relative Entropy** (RE) as the proper ranking function:

$$S[q, p] = -\sum_z q(z) \log \frac{q(z)}{p(z)}$$

- $\Rightarrow$ When information is supplied in the form of constaints on the posterior, we *should* select the posterior that maximizes the Relative Entropy, subject to the constraints. This is the **Method of Maximum (Relative) Entropy** (MRE).

Relative Entropy, KL Divergence and Free Energy

- Note that the Relative Entropy is technically a *functional*, i.e., a function of function(s) (since it is a function of probability distributions). The calculus of functionals is also called **variational calculus**.
  - See Lanczos, The variational principles of mechanics (1961) for a great book on variational calculus. For a summary, see Bishop (2016), app.D.
  - It is customary to use square brackets (like $S[q, p]$) for functionals rather than round brackets, which we use for functions.

- Also note the complimentary relation between the Maximum Relative Entropy method and Probability Theory:
  - PT describes how to *represent* beliefs about events and how to *calculate* beliefs about joint and conditional events.
  - In contrast, the MRE method describes how to *update* beliefs when new information becomes available.

- PT and the MRE method are both needed to describe the theory of optimal information processing. (As we will see below, Bayes rule is a special case of the MRE method.)

- In principle, entropies can always be considered as a *relative* score against a reference distribution. For instance, the score $-\sum_{z_i} q(z_i) \log q(z_i)$ can be interpreted as a score against the uniform distribution, i.e., $-\sum_{z_i} q(z_i) \log q(z_i) \propto -\sum_{z_i} q(z_i) \log \frac{q(z_i)}{\text{Uniform}(z_i)}$. Therefore, the "method of maximum relative entropy" is often simply known as the "method of maximum entropy".

- The negative relative entropy is known as the **Kullback-Leibler** (KL) divergence:

$$\mathrm{KL}[q,p] \triangleq \sum_z q(z) \log \frac{q(z)}{p(z)} \tag{B-1.113}$$

- The Gibbs inequality (proof), is a famous theorem in information theory that states that

$$\mathrm{KL}[q,p] \geq 0$$

  with equality only iff $p = q$.

- The KL divergence can be interpreted as a "distance" between two probability distributions. Note however that the KL divergence is an asymmetric distance measure, i.e., in general $\mathrm{KL}[q,p] \neq \mathrm{KL}[p,q]$.

- We also introduce here the notion of a (variational) **Free Energy** (FE) functional, which is just a generalization of the KL-divergence that allows the prior to be unnormalized. Let $f(z) = Z \cdot p(z)$ be an unnormalized distribution, then the FE is defined as

$$\begin{aligned} F[q,p] &\triangleq \sum_z q(z) \log \frac{q(z)}{f(z)} \\ &= \sum_z q(z) \log \frac{q(z)}{Z \cdot p(z)} \\ &= \underbrace{\sum_z q(z) \log \frac{q(z)}{p(z)}}_{\text{KL divergence} \geq 0} - \log Z \end{aligned}$$

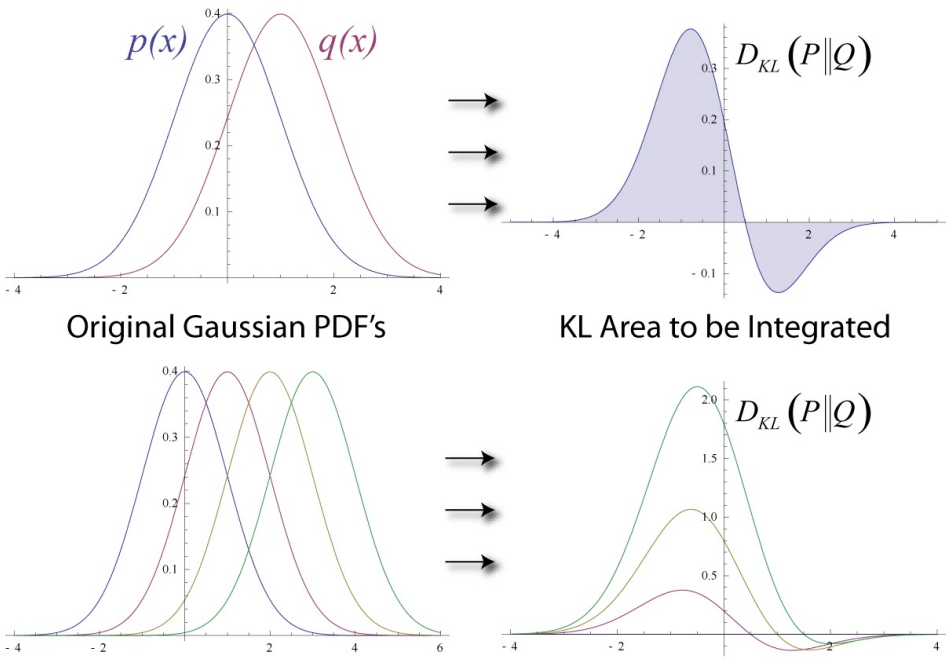
- Since the second term ($\log Z$) is constant, FE minimization (subject to constraints) with respect to q leads to the same posteriors as KL minimization and RE maximization.

- If we are only interested in minimizing FE with respect to q, then we'll write $F[q]$ (rather than $F[q,p]$) fo brevity.

- In this class, we prefer to discuss inference in terms of minimizing Free Energy (subject to the constraints) rather than maximizing Relative Entropy, but note that these two concepts are equivalent.

Example KL divergence for Gaussians

The following picture illustrates the KL-divergence between two Gaussian distributions.



source: By Mundhenk at English Wikipedia, CC BY-SA 3.0, Link

The Free Energy Functional and Variational Bayes

- Let's get back to the issue of doing inference for models with latent variables (such as the GMM).

- Consider a generative model specified by $p(x,z)$ where x and z represent the observed and unobserved variables, respectively.
- Assume further that x has been observed as $x=\hat{x}$ and we are interested in performing inference, i.e., we want to compute the posterior $p(z|x=\hat{x})$ for the latent variables and we want to compute the evidence $p(x=\hat{x})$.
- According to the Method of Maximum Relative Entropy, in order to find the correct posterior $q(x,z)$, we should minimize

$$\mathrm{F}[q] = \sum_{x,z} q(x,z) \log \frac{q(x,z)}{p(x,z)}, \quad \text{subject to data constraint } x=\hat{x}$$

- The data constraint $x=\hat{x}$ fixes posterior $q(x)=\delta(x-\hat{x})$, so the Free Energy evaluates to

$$\begin{aligned} F[q] &= \sum_z \sum_x q(z|x)q(x) \log \frac{q(z|x)q(x)}{p(z|x)p(x)} \\ &= \sum_z \sum_x q(z|x)\delta(x-\hat{x}) \log \frac{q(z|x)\delta(x-\hat{x})}{p(z|x)p(x)} \\ &= \sum_z q(z|\hat{x}) \log \frac{q(z|\hat{x})}{p(z|\hat{x})p(\hat{x})} \\ &= \underbrace{\sum_z q(z|\hat{x}) \log \frac{q(z|\hat{x})}{p(z|\hat{x})}}_{\text{KL divergence } \geq 0} - \underbrace{\log p(\hat{x})}_{\text{log-evidence}} \end{aligned} \tag{B-10.2}$$

- The log-evidence term does not depend on q. Hence, the global minimum $q(z|\hat{x})^* \triangleq \arg\min_q F[q]$ is obtained for

$$q^*(z|\hat{x}) = p(z|\hat{x}),$$

  which is the correct **Bayesian posterior**.
- Furthermore, the minimal free energy

$$F^* \triangleq F[q^*] = -\log p(\hat{x})$$

  equals minus **log-evidence**. (Or, equivalently, the evidence is given by $p(\hat{x}) = \exp(-F[q^*])$.)
- This is an amazing result! Note that FE minimization transforms an inference problem (that involves integration) to an optimization problem! Generally, optimization problems are easier to solve than integration problems.
- Executing inference by minimizing the variational FE functional is called **Variational Bayes** (VB) or variational inference.
- (As an aside), note that Bishop introduces in Eq. B-10.2 an *Evidence Lower BOund* (in modern machine learning literature abbreviated as **ELBO**) $\mathcal{L}(z)$ that equals the *negative* FE. We prefer to discuss variational inference in terms of a free energy (but it is the same story as he discusses with the ELBO).

FE Minimization as Approximate Bayesian Inference

- In the rest of this lesson, we are concerned with how to execute FE minimization (FEM) w.r.t q for the functional

$$F[q] = \sum_z q(z) \log \frac{q(z)}{p(z|x)} - \log p(x)$$

  where x has been observed and z represent all latent variables.
  - To keep the notation uncluttered, in the following we write x rather than $\hat{x}$, and we write simply $q(z)$ (rather than $q(z|x)$) for the posterior.
- Due to restrictions in our optimization algorithm, we are often unable to fully minimize the FE to the global minimum $q^*(z)$, but rather get stuck in a local minimum $\hat{q}(z)$.
- Note that, since $\mathrm{KL}[q(z), p(z|x)] \geq 0$ for any $q(z)$, the FE is always an upperbound on (minus) log-evidence, i.e.,

$$F[q] \geq -\log p(x).$$

- In practice, even if we cannot attain the global minimum, we can still use the local minimum $\hat{q}(z) \approx \arg\min_q F[q]$ to accomplish **approximate Bayesian inference**:

$$\begin{aligned} &\text{posterior: } \hat{q}(z) \approx p(z|x) \\ &\text{evidence: } p(x) \approx \exp(-F[q]) \end{aligned}$$

Constrained FE Minimization

- Generally speaking, it can be very helpful in the FE minimization task to add some additional constraints on $q(z)$ (i.e., in addition to the data constraints).

- Once we add constraints to q (in addition to data constraints), we are no longer guaranteed that the minimum of the (constrained) FE coincides with the solution by Bayes rule (which only takes data as constraints). So again, at best constrained FEM is only an **approximation to Bayes rule**.

- There are three important cases of adding constraints to $q(z)$ that often alleviates the FE minimization task:

1. mean-field factorization

- We constrain the posterior to factorize into a set of *independent* factors, i.e.,

$$q(z) = \prod_{j=1}^{m} q_j(z_j) \,, \tag{B-10.5}$$

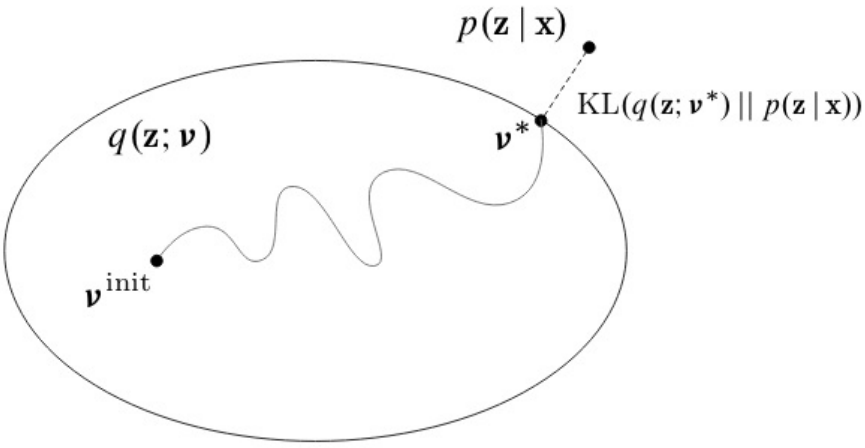
  - Variational inference with mean-field factorization has been worked out in detail as the **Coordinate Ascent Variational Inference** (CAVI) algorithm. See the Optional Slide on CAVI for details.

2. fixed-form parameterization

- We constrain the posterior to be part of a parameterized probability distribution, e.g.,

$$q(z) = \mathcal{N}(z|\mu, \Sigma) \,.$$

  - In this case, the functional minimization problem for $F[q]$ reduces to the minimization of a *function* $F(\mu, \Sigma)$ w.r.t. its parameters. We can often use standand gradient-based optimization methods to minimize the FE.
  - The following image by David Blei illustrates this approach



3. the Expectation-Maximization (EM) algorithm

- We place some constraints both on the prior and posterior for z (also to be discussed in an Optional Slide) that simplifies FE minimization to maximum-likelihood estimation.

Example: FEM for the Gaussian Mixture Model (CAVI Approach)

- Let's get back to the illustrative example at the beginning of this lesson: we want to do density modeling for the Old Faithful data set.

model specification

- We consider a Gaussian Mixture Model, specified by

$$\begin{aligned} p(x, z|\theta) &= p(x|z, \mu, \Lambda)p(z|\pi) \\ &= \prod_{n=1}^{N} \prod_{k=1}^{K} \left(\pi_k \cdot \mathcal{N}\left(x_n|\mu_k, \Lambda_k^{-1}\right)\right)^{z_{nk}} \end{aligned} \tag{B-10.37,38}$$

  with tuning parameters $\theta = \{\pi_k, \mu_k, \Lambda_k\}$.

- Let us introduce some priors for the parameters. We factorize the prior and choose conjugate distributions by

$$p(\theta) = p(\pi, \mu, \Lambda) = p(\pi)p(\mu|\Lambda)p(\Lambda)$$

  with

$$p(\pi) = \text{Dir}(\pi|\alpha_0) = C(\alpha_0) \prod_k \pi_k^{\alpha_0 - 1} \tag{B-10.39}$$

$$p(\mu|\Lambda) = \prod_k \mathcal{N}\left(\mu_k | m_0, (\beta_0 \Lambda_k)^{-1}\right) \tag{B-10.40}$$

$$p(\Lambda) = \prod_k \mathcal{W}\left(\Lambda_k | W_0, \nu_0\right) \tag{B-10.40}$$

where $\mathcal{W}(\cdot)$ is a Wishart distribution (i.e., a multi-dimensional Gamma distribution).

- The full generative model is now specified by

$$p(x, z, \pi, \mu, \Lambda) = p(x|z, \mu, \Lambda) p(z|\pi) p(\pi) p(\mu|\Lambda) p(\Lambda) \tag{B-10.41}$$

with hyperparameters $\{\alpha_0, m_0, \beta_0, W_0, \nu_0\}$.

inference

- Assume that we have observed $D = \{x_1, x_2, \ldots, x_N\}$ and are interested to infer the posterior distribution for the tuning parameters:

$$p(\theta|D) = p(\pi, \mu, \Lambda|D)$$

- The (perfect) Bayesian solution is

$$p(\theta|D) = \frac{p(x=D, \theta)}{p(x=D)} = \frac{\sum_z p(x=D, z, \pi, \mu, \Lambda)}{\sum_z \sum_\pi \iint p(x=D, z, \pi, \mu, \Lambda)\, \mathrm{d}\mu \mathrm{d}\Lambda}$$

but this is intractable (See Blei (2017), p861, eqs. 8 and 9).

- Alternatively, we can use **FE minimization with factorization constraint**

$$q(\theta) = q(z) \cdot q(\pi, \mu, \Lambda) \tag{B-10.42}$$

on the posterior. For the specified model, this extra constaint leads to FE minimization wrt the hyperparameters, i.e., we need to minimize the function $F(\alpha_0, m_0, \beta_0, W_0, \nu_0)$.

- Bishop shows that the equations for the optimal solutions (Eq. B-10.9) are analytically solvable for the GMM as specified above, leading to the following variational update equations (for $k = 1, \ldots, K$):

$$\alpha_k = \alpha_0 + N_k \tag{B-10.58}$$

$$\beta_k = \beta_0 + N_k \tag{B-10.60}$$

$$m_k = \frac{1}{\beta_k}\left(\beta_0 m_0 + N_k \bar{x}_k\right) \tag{B-10.61}$$

$$W_k^{-1} = W_0^{-1} + N_k S_k + \frac{\beta_0 N_k}{\beta_0 + N_k}\left(x_k - m_0\right)\left(x_k - m_0\right)^T \tag{B-10.62}$$

$$\nu_k = \nu_0 + N_k \tag{B-10.63}$$

where we used

$$\log \rho_{nk} = \mathbb{E}\left[\log \pi_k\right] + \frac{1}{2}\mathbb{E}\left[\log |\Lambda_k|\right] - \frac{D}{2}\log(2\pi) - \frac{1}{2}\mathbb{E}\left[(x_k - \mu_k)^T \Lambda_k (x_k - \mu_k)\right] \tag{B-10.46}$$

$$r_{nk} = \frac{\rho_{nk}}{\sum_{j=1}^K \rho_{nj}} \tag{B-10.49}$$

$$N_k = \sum_{n=1}^N r_{nk} x_n \tag{B-10.51}$$

$$\bar{x}_k = \frac{1}{N_k} \sum_{n=1}^N r_{nk} x_n \tag{B-10.52}$$

$$S_k = \frac{1}{N_k} \sum_{n=1}^N r_{nk} \left(x_n - \bar{x}_k\right)\left(x_n - \bar{x}_k\right)^T \tag{B-10.53}$$

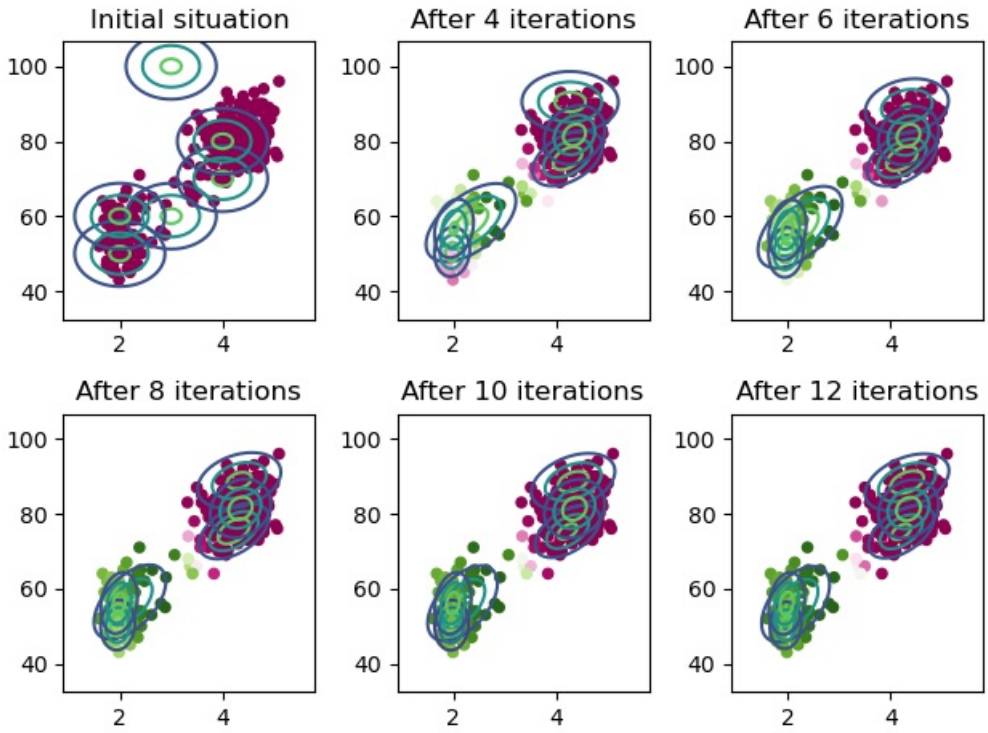
- Exam guide: Working out FE minimization for the GMM to these update equations (eqs B-10.58 through B-10.63) is not something that you need to reproduce without assistance at the exam. Rather, the essence is that *it is possible* to arrive at closed-form variational update equations for the GMM. You should understand though how FEM works conceptually and in principle be able to derive variational update equations for very simple models that do not involve clever mathematical tricks.

Code Example: FEM for GMM on Old Faithfull data set

- Below we exemplify training of a Gaussian Mixture Model on the Old Faithful data set by Free Energy Minimization, using the constraints as specified above, e.g., (B-10.42).

In [1]:
```
using DataFrames, CSV, LinearAlgebra, PDMats, SpecialFunctions
include("scripts/gmm_plot.jl") # Holds plotting function
old_faithful = CSV.read("datasets/old_faithful.csv",DataFrame);
X = convert(Matrix{Float64}, [old_faithful[!,1] old_faithful[!,2]]');#data matrix
N = size(X, 2) #number of observations
```

```
K = 6

function sufficientStatistics(X,r,k::Int) #function to compute sufficient statistics
    N_k = sum(r[k,:])
    hat_x_k = sum([r[k,n]*X[:,n] for n in 1:N]) ./ N_k
    S_k = sum([r[k,n]*(X[:,n]-hat_x_k)*(X[:,n]-hat_x_k)' for n in 1:N]) ./ N_k
    return N_k, hat_x_k, S_k
end

function updateMeanPrecisionPi(m_0,β_0,W_0,ν_0,α_0,r) #variational maximisation function
    m = Array{Float64}(undef,2,K) #mean of the clusters
    β = Array{Float64}(undef,K) #precision scaling for Gausian distribution
    W = Array{Float64}(undef,2,2,K) #precision prior for Wishart distributions
    ν = Array{Float64}(undef,K) #degrees of freedom parameter for Wishart distribution
    α = Array{Float64}(undef,K) #Dirichlet distribution parameter
    for k=1:K
        sst = sufficientStatistics(X,r,k)
        α[k] = α_0[k] + sst[1]; β[k] = β_0[k] + sst[1]; ν[k] = ν_0[k] .+ sst[1]
        m[:,k] = (1/β[k])*(β_0[k].*m_0[:,k] + sst[1].*sst[2])
        W[:,:,k] = inv(inv(W_0[:,:,k])+sst[3]*sst[1] + ((β_0[k]*sst[1])/(β_0[k]+sst[1])).*(sst[2]-m_0[:,k])*(sst[2]-m
    end
    return m,β,W,ν,α
end

function updateR(Λ,m,α,ν,β) #variational expectation function
    r = Array{Float64}(undef,K,N) #responsibilities
    hat_π = Array{Float64}(undef,K)
    hat_Λ = Array{Float64}(undef,K)
    for k=1:K
        hat_Λ[k] = 1/2*(2*log(2)+logdet(Λ[:,:,k])+digamma(ν[k]/2)+digamma((ν[k]-1)/2))
        hat_π[k] = exp(digamma(α[k])-digamma(sum(α)))
        for n=1:N
           r[k,n] = hat_π[k]*exp(-hat_Λ[k]-1/β[k] - (ν[k]/2)*(X[:,n]-m[:,k])'*Λ[:,:,k]*(X[:,n]-m[:,k]))
        end
    end
    for n=1:N
        r[:,n] = r[:,n]./ sum(r[:,n]) #normalize to ensure r represents probabilities
    end
    return r
end

max_iter = 15
#store the inference results in these vectors
ν = fill!(Array{Float64}(undef,K,max_iter),3)
β = fill!(Array{Float64}(undef,K,max_iter),1.0)
α = fill!(Array{Float64}(undef,K,max_iter),0.01)
R = Array{Float64}(undef,K,N,max_iter)
M = Array{Float64}(undef,2,K,max_iter)
Λ = Array{Float64}(undef,2,2,K,max_iter)
clusters_vb = Array{Distribution}(undef,K,max_iter) #clusters to be plotted
#initialize prior distribution parameters
M[:,:,1] = [[3.0; 60.0];[4.0; 70.0];[2.0; 50.0];[2.0; 60.0];[3.0; 100.0];[4.0; 80.0]]
for k=1:K
    Λ[:,:,k,1] = [1.0 0;0 0.01]
    R[k,:,1] = 1/(K)*ones(N)
    clusters_vb[k,1] = MvNormal(M[:,k,1],PDMats.PDMat(convert(Matrix,Hermitian(inv(ν[1,1].*Λ[:,:,k,1])))))
end
#variational inference
for i=1:max_iter-1
    #variational expectation
    R[:,:,i+1] = updateR(Λ[:,:,:,i],M[:,:,i],α[:,i],ν[:,i],β[:,i])
    #variational minimisation
    M[:,:,i+1],β[:,i+1],Λ[:,:,:,i+1],ν[:,i+1],α[:,i+1] = updateMeanPrecisionPi(M[:,:,i],β[:,i],Λ[:,:,:,i],ν[:,i],α[:,
    for k=1:K
        clusters_vb[k,i+1] = MvNormal(M[:,k,i+1],PDMats.PDMat(convert(Matrix,Hermitian(inv(ν[k,i+1].*Λ[:,:,k,i+1])))))
    end
end


subplot(2,3,1); plotGMM(X, clusters_vb[:,1], R[:,:,1]); title("Initial situation")
for i=2:6
    subplot(2,3,i)
    plotGMM(X, clusters_vb[:,i*2], R[:,:,i*2]); title("After $(i*2) iterations")
end
PyPlot.tight_layout();
```



The generated figure looks much like Figure 10.6 in Bishop. The plots show results for Variational Bayesian mixture of K = 6 Gaussians applied to the Old Faithful data set, in which the ellipses denote the one standard-deviation density contours for each of the components, and the color coding of the data points reflects the "soft" class label assignments. Components whose expected mixing coefficient are numerically indistinguishable from zero are not plotted. Note that this procedure learns the number of classes (two), learns the class label for each observation, and learns the mean and variance for the Gaussian data distributions.

Interesting Decompositions of the Free Energy Functional

- The FE functional can be decomposed in various interesting ways, making use of $p(x,z) = p(z|x)p(x) = p(x|z)p(z)$

$$\mathrm{F}[q] = \underbrace{-\sum_z q(z)\log p(x,z)}_{\text{energy}} \underbrace{-\sum_z q(z)\log\frac{1}{q(z)}}_{\text{entropy}} \tag{EE}$$

$$= \underbrace{\sum_z q(z)\log\frac{q(z)}{p(z|x)}}_{\text{KL divergence}\geq 0} \underbrace{-\log p(x)}_{\text{log-evidence}} \tag{DE}$$

$$= \underbrace{\sum_z q(z)\log\frac{q(z)}{p(z)}}_{\text{complexity}} \underbrace{-\sum_z q(z)\log p(x|z)}_{\text{accuracy}} \tag{CA}$$

- These decompositions are very insightful (we will revisit them later) and we will label them respectively as *energy-entropy* (EE), *divergence-evidence* (DE), and *complexity-accuracy* (CA) decompositions.
- In the Bayesian Machine Learning lecture, we discussed the CA decomposition of Bayesian model evidence to support the interpretation of evidence as a model performance criterion. Here, we recognize that FE allows a similar CA decomposition: minimizing FE increases data fit and decreases model complexity. Hence, FE is a good model performance criterion.
- The CA decomposition makes use of the prior $p(z)$ and likelihood $p(x|z)$, both of which are selected by the engineer, so the FE can be evaluated with this decomposition!
- The DE decomposition restates what we derived earlier, namely that the FE is an upperbound on the (negative) log-evidence. The bound is the KL-divergence between the variational posterior $q(z)$ and the (perfect) Bayesian posterior $p(z|x)$. Global minimization of FE with only data constraints drives the KL-divergence to zero and results to perfect Bayesian inference.
- The EE decomposition provides a link to the second law of thermodynamics: Minimizing FE leads to entropy maximization, subject to constraints, where in this case the constraints are imposed by the postulated generative model.

Summary

- Latent variable models (LVM) contain a set of unobserved variables whose size grows with the number of observations.
- LVMs can model more complex phenomena than fully observed models, but inference in LVM models is usually not analytically solvable.

- The Free Energy (FE) functional transforms Bayesian inference computations (very large summations or integrals) to an optimization problem.

- Inference by minimizing FE, also known as variational inference, is fully consistent with the "Method of Maximum Relative Entropy", which is by design the rational way to update beliefs from priors to posteriors when new information becomes available. Thus, FE mimimization is the "correct" inference procedure that generalizes Bayes rule.

- In general, global FE minimization is an unsolved problem and finding good local minima is at the heart of current Bayesian technology research. Three simplifying constraints on the posterior $q(z)$ in the FE functional are currently popular in practical settings:
  - mean-field assumptions
  - assuming a parametric from for q
  - EM algorithm

- These constraints often makes FE minimization implementable at the price of obtaining approximately Bayesian inference results.

- The back-ends of Probabilistic Programming Languages (such as ReactiveMP.jl) often contain lots of methods for automating constrained FE minimization.

OPTIONAL SLIDES

FE Minimization with Mean-field Factorization Constraints: the CAVI Approach

- Let's work out FE minimization with additional mean-field constraints (=full factorization) constraints:

$$q(z) = \prod_{j=1}^{m} q_j(z_j) \,.$$

  - In other words, the posteriors for z_j are all considered independent. This is a strong constraint but leads often to good solutions.
- Given the mean-field constraints, it is possible to derive the following expression for the optimal solutions $q_j^*(z_j)$, for $j = 1, \ldots, m$:

$$\boxed{\begin{aligned}\log q_j^*(z_j) &\propto \mathrm{E}_{q^*_{-j}}[\log p(x,z)] \\ &= \underbrace{\sum_{z_{-j}} q^*_{-j}(z_{-j}) \underbrace{\log p(x,z)}_{\text{"field"}}}_{\text{"mean field"}}\end{aligned}} \tag{B-10.9}$$

where we defined $q^*_{-j}(z_{-j}) \triangleq q_1^*(z_1) q_2^*(z_2) \cdots q_{j-1}^*(z_{j-1}) q_{j+1}^*(z_{j+1}) \cdots q_m^*(z_m)$.

- **Proof** (from Blei, 2017): We first rewrite the FE as a function of $q_j(z_j)$ only:

$$F[q_j] = \mathbb{E}_{q_j}\Big[\mathbb{E}_{q_{-j}}\Big[\log p(x, z_j, z_{-j})\Big]\Big] - \mathbb{E}_{q_j}\Big[\log q_j(z_j)\Big] + \texttt{const.}\,,$$

  where the constant holds all terms that do not depend on z_j. This expression can be written as

$$F[q_j] = \sum_{z_j} q_j(z_j) \log \frac{q_j(z_j)}{\exp\Big(\mathbb{E}_{q_{-j}}\Big[\log p(x, z_j, z_{-j})\Big]\Big)}$$

  which is a KL-divergence that is minimized by Eq. B-10.9. (end proof)

- This is not yet a full solution to the FE minimization task since the solution $q_j^*(z_j)$ depends on expectations that involve other solutions $q^*_{i\neq j}(z_{i\neq j})$, and each of these other solutions $q^*_{i\neq j}(z_{i\neq j})$ depends on an expection that involves $q_j^*(z_j)$.

- In practice, we solve this chicken-and-egg problem by an iterative approach: we first initialize all $q_j(z_j)$ (for $j = 1, \ldots, m$) to an appropriate initial distribution and then cycle through the factors in turn by solving eq.B-10.9 and update $q^*_{-j}(z_{-j})$ with the latest estimates. (See Blei, 2017, Algorithm 1, p864).

- This algorithm for approximating Bayesian inference is known **Coordinate Ascent Variational Inference** (CAVI).

FE Minimization by the Expectation-Maximization (EM) Algorithm

- The EM algorithm is a special case of FE minimization that focusses on Maximum-Likelihood estimation for models with latent variables.
- Consider a model

$$p(x, z, \theta)$$

  with observations $x = \{x_n\}$, latent variables $z = \{z_n\}$ and parameters θ.

- We can write the following FE functional for this model:

$$F[q] = \sum_{z} \sum_{\theta} q(z, \theta) \log \frac{q(z, \theta)}{p(x, z, \theta)}$$

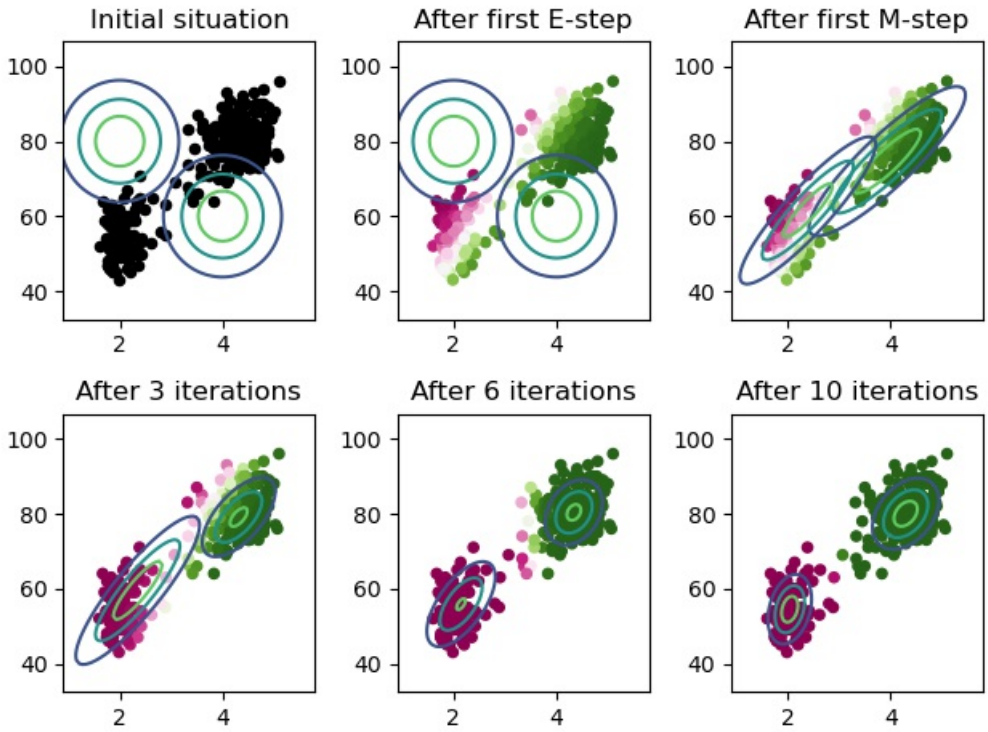
- The EM algorithm makes the following simplifying assumptions:
  1. The prior for the parameters is uninformative (uniform). This implies that
  $$p(x, z, \theta) = p(x, z|\theta)p(\theta) \propto p(x, z|\theta)$$
  2. A factorization constraint
  $$q(z, \theta) = q(z)q(\theta)$$
  3. The posterior for the parameters is a delta function:
  $$q(\theta) = \delta(\theta - \hat{\theta})$$
- Basically, these three assumptions turn FE minimization into maximum likelihood estimation for the parameters θ and the FE simplifies to
$$F[q, \theta] = \sum_z q(z) \log \frac{q(z)}{p(x, z|\theta)}$$
- The EM algorithm minimizes this FE by iterating (iteration counter: i) over
$$\boxed{\begin{aligned} \mathcal{L}^{(i)}(\theta) &= \sum_z \overbrace{p(z|x, \theta^{(i-1)})}^{q^{(i)}(z)} \log p(x, z|\theta) && \text{the E-step} \\ \theta^{(i)} &= \arg\max_\theta \mathcal{L}^{(i)}(\theta) && \text{the M-step} \end{aligned}}$$
- These choices are optimal for the given FE functional. In order to see this, consider the two decompositions
$$F[q, \theta] = \underbrace{-\sum_z q(z) \log p(x, z|\theta)}_{\text{energy}} \underbrace{-\sum_z q(z) \log \frac{1}{q(z)}}_{\text{entropy}} \tag{EE}$$
$$= \underbrace{\sum_z q(z) \log \frac{q(z)}{p(z|x, \theta)}}_{\text{divergence}} \underbrace{- \log p(x|\theta)}_{\text{log-likelihood}} \tag{DE}$$
- The DE decomposition shows that the FE is minimized for the choice $q(z) := p(z|x, \theta)$. Also, for this choice, the FE equals the (negative) log-evidence (, which is this case simplifies to the log-likelihood).
- The EE decomposition shows that the FE is minimized wrt θ by minimizing the energy term. The energy term is computed in the E-step and optimized in the M-step.
  - Note that in the EM literature, the energy term is often called the *expected complete-data log-likelihood*.)
- In order to execute the EM algorithm, it is assumed that we can analytically execute the E- and M-steps. For a large set of models (including models whose distributions belong to the exponential family of distributions), this is indeed the case and hence the large popularity of the EM algorithm.
- The EM algorihm imposes rather severe assumptions on the FE (basically approximating Bayesian inference by maximum likelihood estimation). Over the past few years, the rise of Probabilistic Programming languages has dramatically increased the range of models for which the parameters can by estimated autmatically by (approximate) Bayesian inference, so the popularity of EM is slowly waning. (More on this in the Probabilistic Programming lessons).
- Bishop (2006) works out EM for the GMM in section 9.2.

Code Example: EM-algorithm for the GMM on the Old-Faithful data set

We'll perform clustering on the data set from the illustrative example by fitting a GMM consisting of two Gaussians using the EM algorithm.

In [2]:
```
using DataFrames, CSV, LinearAlgebra
include("scripts/gmm_plot.jl") # Holds plotting function
old_faithful = CSV.read("datasets/old_faithful.csv", DataFrame);

X =  Array(Matrix{Float64}(old_faithful)')
N = size(X, 2)

# Initialize the GMM. We assume 2 clusters.
clusters = [MvNormal([4.;60.], [.5 0;0 10^2]);
            MvNormal([2.;80.], [.5 0;0 10^2])];
π_hat = [0.5; 0.5]                    # Mixing weights
γ = fill!(Matrix{Float64}(undef,2,N), NaN)  # Responsibilities (row per cluster)

# Define functions for updating the parameters and responsibilities
function updateResponsibilities!(X, clusters, π_hat, γ)
    # Expectation step: update γ
    norm = [pdf(clusters[1], X) pdf(clusters[2], X)] * π_hat
    γ[1,:] = (π_hat[1] * pdf(clusters[1],X) ./ norm)'
    γ[2,:] = 1 .- γ[1,:]
end
function updateParameters!(X, clusters, π_hat, γ)
```

```
    # Maximization step: update π_hat and clusters using ML estimation
    m = sum(γ, dims=2)
    π_hat = m / N
    μ_hat = (X * γ') ./ m'
    for k=1:2
        Z = (X .- μ_hat[:,k])
        Σ_k = Symmetric(((Z .* (γ[k,:])') * Z') / m[k])
        clusters[k] = MvNormal(μ_hat[:,k], convert(Matrix, Σ_k))
    end
end

# Execute the algorithm: iteratively update parameters and responsibilities
subplot(2,3,1); plotGMM(X, clusters, γ); title("Initial situation")
updateResponsibilities!(X, clusters, π_hat, γ)
subplot(2,3,2); plotGMM(X, clusters, γ); title("After first E-step")
updateParameters!(X, clusters, π_hat, γ)
subplot(2,3,3); plotGMM(X, clusters, γ); title("After first M-step")
iter_counter = 1
for i=1:3
    for j=1:i+1
        updateResponsibilities!(X, clusters, π_hat, γ)
        updateParameters!(X, clusters, π_hat, γ)
        iter_counter += 1
    end
    subplot(2,3,3+i);
    plotGMM(X, clusters, γ);
    title("After $(iter_counter) iterations")
end
PyPlot.tight_layout()
```



Note that you can step through the interactive demo yourself by running this script in julia. You can run a script in julia by

```
julia> include("path/to/script-name.jl")
```

Message Passing for Free Energy Minimization

- The Sum-Product (SP) update rule implements perfect Bayesian inference.
- Sometimes, the SP update rule is not analytically solvable.
- Fortunately, for many well-known Bayesian approximation methods, a message passing update rule can be created, e.g. Variational Message Passing (VMP) for variational inference.
- In general, all of these message passing algorithms can be interpreted as minimization of a constrained free energy (e.g., see Senoz et al. (2021), and hence these message passing schemes comply with Caticha's Method of Maximum Relative Entropy, which, as discussed in the variational Bayes lesson is the proper way for updating beliefs.
- Different message passing updates rules can be combined to get a hybrid inference method in one model.

The Local Free Energy in a Factor Graph

- Consider an edge x_j in a Forney-style factor graph for a generative model $p(x) = p(x_1, x_2, \ldots, x_N)$.
- Assume that the graph structure (factorization) is specified by

$$p(x) = \prod_{a=1}^{M} p_a(x_a)$$

  where a is a set of indices.
- Also, we assume a mean-field approximation for the posterior:

$$q(x) = \prod_{i=1}^{N} q_i(x_i)$$

  and consequently a corresponding free energy functional

$$\begin{aligned} F[q] &= \sum_x q(x) \log \frac{q(x)}{p(x)} \\ &= \sum_i \sum_{x_i} \left(\prod_{i=1}^{N} q_i(x_i) \right) \log \frac{\prod_{i=1}^{N} q_i(x_i)}{\prod_{a=1}^{M} p_a(x_a)} \end{aligned}$$

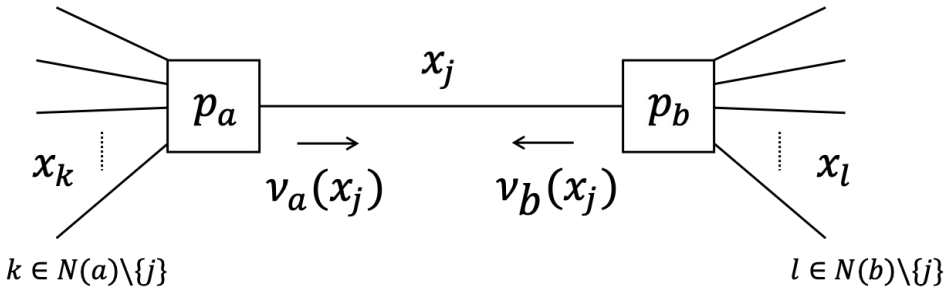
- With these assumptions, it can be shown that the FE evaluates to (exercise)

$$F[q] = \underbrace{\sum_{a=1}^{M} \sum_{x_a} \left(\prod_{j \in N(a)} q_j(x_j) \cdot (-\log p_a(x_a)) \right)}_{\text{node energy } U[p_a]} \underbrace{- \sum_{i=1}^{N} \sum_{x_i} q_i(x_i) \log \frac{1}{q_i(x_i)}}_{\text{edge entropy } H[q_i]}$$

- In words, the FE decomposes into a sum of (expected) energies for the nodes minus the entropies on the edges.

Variational Message Passing

- Let us now consider the local free energy that is associated with edge corresponding to x_j.



- Apparently (see previous slide), there are three contributions to the free energy for x_j:
  - one entropy term for the edge x_j
  - two energy terms: one for each node that attaches to x_j (in the figure: nodes p_a and p_b)
- The local free energy for x_j can be written as (exercise)

$$F[q_j] \propto \sum_{x_j} q(x_j) \log \frac{q_j(x_j)}{\nu_a(x_j) \cdot \nu_b(x_j)}$$

  where

$$\begin{aligned} \nu_a(x_j) &\propto \exp\left(\mathbb{E}_{q_k}[\log p_a(x_a)] \right) \\ \nu_b(x_j) &\propto \exp\left(\mathbb{E}_{q_l}[\log p_b(x_b)] \right) \end{aligned}$$

  and $\mathbb{E}_{q_k}[\cdot]$ is an expectation w.r.t. all $q(x_k)$ with $k \in N(a) \setminus j$.
- $\nu_a(x_j)$ and $\nu_b(x_j)$ can be locally computed in nodes a and b respectively and can be interpreted as colliding messages over edge x_j.
- Local free energy minimization is achieved by setting

$$q_j(x_j) \propto \nu_a(x_j) \cdot \nu_b(x_j)$$

- Note that message $\nu_a(x_j)$ depends on posterior beliefs over incoming edges (k) for node a, and in turn, the message from node a towards edge x_k depends on the belief $q_j(x_j)$. I.o.w., direct mutual dependencies exist between posterior beliefs over edges that attach to the same node.
- These considerations lead to the Variational Message Passing procedure, which is an iterative free energy minimization procedure that can be executed completely through locally computable messages.

- Procedure VMP, see Dauwels (2007), section 3

    1. Initialize all messages q and ν, e.g., $q(\cdot) \propto 1$ and $\nu(\cdot) \propto 1$.
    2. Select an edge z_k in the factor graph of $f(z_1, \ldots, z_m)$.
    3. Compute the two messages $\vec{\nu}(z_k)$ and $\overleftarrow{\nu}(z_k)$ by applying the following generic rule:

    $$\vec{\nu}(y) \propto \exp\left(\mathbb{E}_q\left[\log g(x_1, \ldots, x_n, y)\right]\right)$$

    4. Compute the marginal $q(z_k)$

    $$q(z_k) \propto \vec{\nu}(z_k)\overleftarrow{\nu}(z_k)$$

    and send it to the two nodes connected to the edge x_k.
    5. Iterate 2–4 until convergence.

Dynamic Models

Preliminaries

- Goal
    - Introduction to dynamic (=temporal) Latent Variable Models, including the Hidden Markov Model and Kalman filter.
- Materials
    - Mandatory
        - These lecture notes
    - Optional
        - Bishop pp.605-615 on Hidden Markov Models
        - Bishop pp.635-641 on Kalman filters
        - Faragher (2012), Understanding the Basis of the Kalman Filter
        - Minka (1999), From Hidden Markov Models to Linear Dynamical Systems

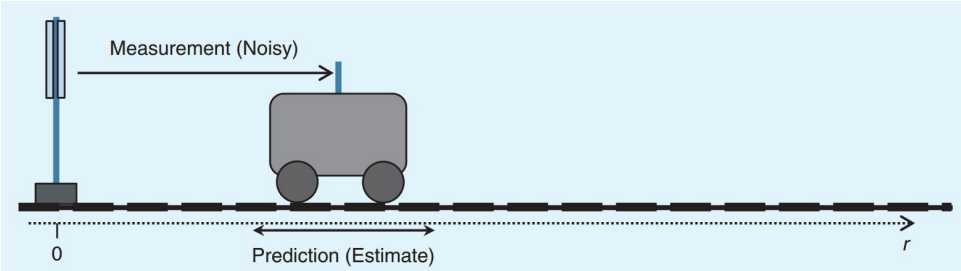
Example Problem

- We consider a one-dimensional cart position tracking problem, see Faragher 2012.
- The hidden states are the position z_t and velocity $\dot{z}_t$. We can apply an external acceleration/breaking force u_t. (Noisy) observations are represented by x_t.
- The equations of motions are given by

$$\begin{bmatrix} z_t \\ \dot{z}_t \end{bmatrix} = \begin{bmatrix} 1 & \Delta t \\ 0 & 1 \end{bmatrix} \begin{bmatrix} z_{t-1} \\ \dot{z}_{t-1} \end{bmatrix} + \begin{bmatrix} (\Delta t)^2/2 \\ \Delta t \end{bmatrix} u_t + \mathcal{N}(0, \Sigma_z) \tag{3}$$

$$x_t = \begin{bmatrix} z_t \\ \dot{z}_t \end{bmatrix} + \mathcal{N}(0, \Sigma_x) \tag{4}$$

- Task: Infer the position z_t after 10 time steps. (Solution later in this lesson).



Dynamical Models

- In this lesson, we consider models where the sequence order of observations matters.
- Consider the *ordered* observation sequence $x^T \triangleq (x_1, x_2, \ldots, x_T)$.
    - (For brevity, in this lesson we use the notation x_t^T to denote $(x_t, x_{t+1}, \ldots, x_T)$ and drop the subscript if $t = 1$, so $x^T = x_1^T = (x_1, x_2, \ldots, x_T)$).
- We wish to develop a generative model

$$p(x^T)$$

that 'explains' the time series x^T.

- We cannot use the IID assumption $p(x^T) = \prod_t p(x_t)$. In general, we *can* use the **chain rule** (a.k.a. **the general product rule**)

$$\begin{aligned} p(x^T) &= p(x_T|x^{T-1})\,p(x^{T-1}) \\ &= p(x_T|x^{T-1})\,p(x_{T-1}|x^{T-2})\cdots p(x_2|x_1)\,p(x_1) \\ &= p(x_1)\prod_{t=2}^{T} p(x_t \mid x^{t-1}) \end{aligned}$$

- Generally, we will want to limit the depth of dependencies on previous observations. For example, a Kth-order linear **Auto-Regressive** (AR) model that is given by

$$p(x_t \mid x^{t-1}) = \mathcal{N}\left(x_t \,\middle|\, \sum_{k=1}^{K} a_k x_{t-k}, \sigma^2\right)$$

  limits the dependencies to the past K samples.

State-space Models

- A limitation of AR models is that they need a lot of parameters in order to create a flexible model. E.g., if x_t is an M-dimensional discrete variable, then a Kth-order AR model will have about M^K parameters.

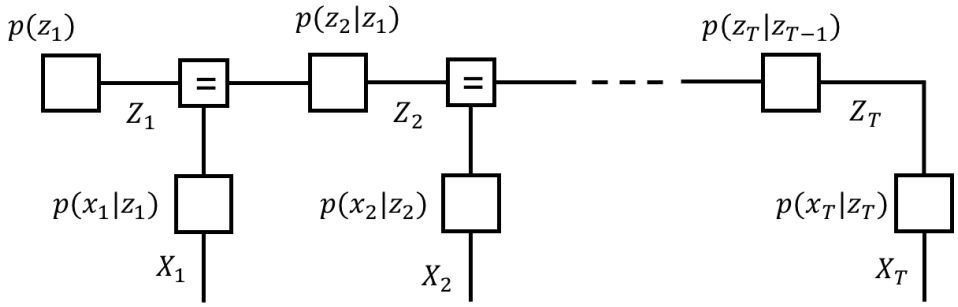
- Similar to our work on Gaussian Mixture models, we can create a flexible dynamic system by introducing *latent* (unobserved) variables $z^T \triangleq (z_1, z_2, \ldots, z_T)$ (one z_t for each observation x_t). In dynamic systems, z_t are called *state variables*.

- A **state space model** is a particular latent variable dynamical model defined by

$$p(x^T, z^T) = \underbrace{p(z_1)}_{\text{initial state}} \prod_{t=2}^{T} \underbrace{p(z_t \mid z_{t-1})}_{\text{state transitions}} \prod_{t=1}^{T} \underbrace{p(x_t \mid z_t)}_{\text{observations}}$$

  - The condition $p(z_t \mid z^{t-1}) = p(z_t \mid z_{t-1})$ is called a 1st-order Markov condition.

- The Forney-style factor graph for a state-space model:



Hidden Markov Models and Linear Dynamical Systems

- A **Hidden Markov Model** (HMM) is a specific state-space model with discrete-valued state variables z_t.

- Typically, z_t is a K-dimensional one-hot coded latent 'class indicator' with transition probabilities $a_{jk} \triangleq p(z_{tk} = 1 \mid z_{t-1,j} = 1)$, or equivalently,

$$p(z_t|z_{t-1}) = \prod_{k=1}^{K}\prod_{j=1}^{K} a_{jk}^{z_{t-1,j} z_{tk}}$$

  which is usually accompanied by an initial state distribution $p(z_{1k} = 1) = \pi_k$.

- The classical HMM has also discrete-valued observations but in pratice any (probabilistic) observation model $p(x_t|z_t)$ may be coupled to the hidden Markov chain.

- Another well-known state-space model with continuous-valued state variables z_t is the **(Linear) Gaussian Dynamical System** (LGDS), which is defined as

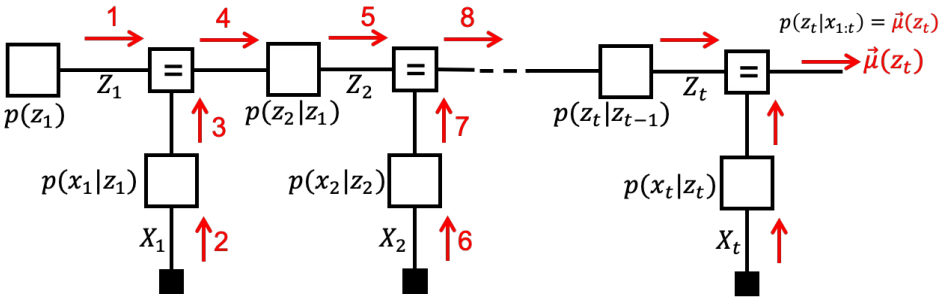
$$\begin{aligned} p(z_t \mid z_{t-1}) &= \mathcal{N}(\,Az_{t-1}\,,\,\Sigma_z\,) \\ p(x_t \mid z_t) &= \mathcal{N}(\,Cz_t\,,\,\Sigma_x\,) \\ p(z_1) &= \mathcal{N}(\,\mu_1\,,\,\Sigma_1\,) \end{aligned}$$

- Note that the joint distribution over all states and observations $\{(x_1, z_1), \ldots, (x_t, z_t)\}$ is a (large-dimensional) Gaussian distribution. This means that, in principle, every inference problem on the LGDS model also leads to a Gaussian distribution.

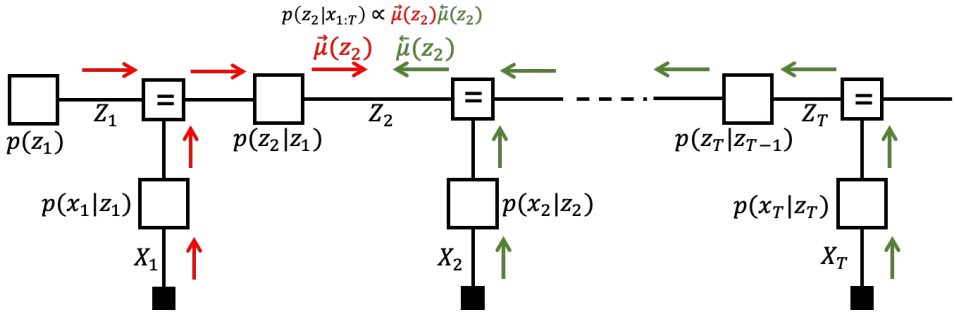
- HMM's and LGDS's (and variants thereof) are at the basis of a wide range of complex information processing systems, such as speech and language recognition, robotics and automatic car navigation, and even processing of DNA sequences.

Common Signal Processing Tasks as Message Passing-based Inference

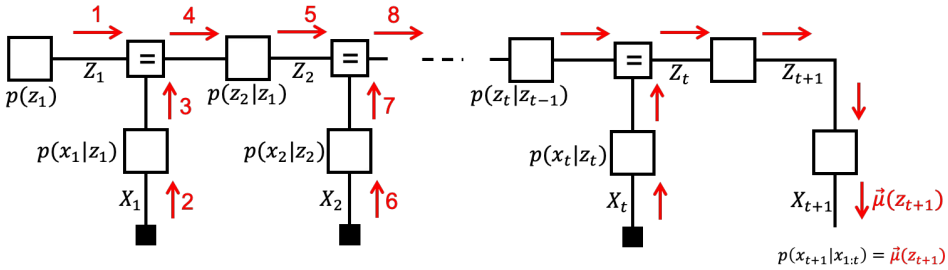
- As we have seen, inference tasks in linear Gaussian state space models can be analytically solved.
- However, these derivations quickly become cumbersome and prone to errors.
- Alternatively, we could specify the generative model in a (Forney-style) factor graph and use automated message passing to infer the posterior over the hidden variables. Here follows some examples.
- **Filtering**, a.k.a. state estimation: estimation of a state (at time step t), based on past and current (at t) observations.



- **Smoothing**: estimation of a state based on both past and future observations. Needs backward messages from the future.



- **Prediction**: estimation of future state or observation based only on observations of the past.



Kalman Filtering

- Technically, a **Kalman filter** is the solution to the recursive estimation (inference) of the hidden state z_t based on past observations in an LGDS, i.e., Kalman filtering solves the problem $p(z_t \mid x^t)$ based on the previous estimate $p(z_{t-1} \mid x^{t-1})$ and a new observation x_t (in the context of the given model specification of course).
- Let's infer the Kalman filter for a scalar linear Gaussian dynamical system:

$$p(z_t \mid z_{t-1}) = \mathcal{N}(z_t \mid az_{t-1}, \sigma_z^2) \quad \text{(state transition)}$$
$$p(x_t \mid z_t) = \mathcal{N}(x_t \mid cz_t, \sigma_x^2) \quad \text{(observation)}$$

- Kalman filtering comprises inferring $p(z_t \mid x^t)$ from a given prior estimate $p(z_{t-1} \mid x^{t-1})$ (available after the previous time step) and a new observation x_t. Let us assume that

$$p(z_{t-1} \mid x^{t-1}) = \mathcal{N}(z_{t-1} \mid \mu_{t-1}, \sigma_{t-1}^2) \quad \text{(prior)}$$

- Note that everything is Gaussian, so it is *in principle* possible to execute inference problems analytically and the result will be a Gaussian posterior:
  - (In the following derivation we make use of the renormalization equality $\mathcal{N}(x \mid cz, \sigma^2) = \frac{1}{c}\mathcal{N}\left(z \mid \frac{x}{c}, \left(\frac{\sigma}{c}\right)^2\right)$).

$$
\begin{aligned}
\underbrace{p(z_t \mid x^t)}_{\text{posterior}} &= p(z_t \mid x_t, x^{t-1}) \propto p(x_t, z_t \mid x^{t-1}) \\
&\propto p(x_t \mid z_t)\, p(z_t \mid x^{t-1}) \\
&= p(x_t \mid z_t) \sum_{z_{t-1}} p(z_t, z_{t-1} \mid x^{t-1}) \\
&= \underbrace{p(x_t \mid z_t)}_{\text{observation}} \sum_{z_{t-1}} \underbrace{p(z_t \mid z_{t-1})}_{\text{state transition}} \underbrace{p(z_{t-1} \mid x^{t-1})}_{\text{prior}} \\
&= \mathcal{N}(x_t \mid cz_t, \sigma_x^2) \sum_{z_{t-1}} \mathcal{N}(z_t \mid az_{t-1}, \sigma_z^2)\mathcal{N}(z_{t-1} \mid \mu_{t-1}, \sigma_{t-1}^2) \\
&= \frac{1}{c}\mathcal{N}\left(z_t \mid \frac{x_t}{c}, \left(\frac{\sigma_x}{c}\right)^2\right) \sum_{z_{t-1}} \underbrace{\frac{1}{a}\mathcal{N}\left(z_{t-1} \mid \frac{z_t}{a}, \left(\frac{\sigma_z}{a}\right)^2\right) \mathcal{N}(z_{t-1} \mid \mu_{t-1}, \sigma_{t-1}^2)}_{\text{use Gaussian multiplication formula SRG-6}} \\
&\propto \underbrace{\mathcal{N}\left(z_t \mid \frac{x_t}{c}, \left(\frac{\sigma_x}{c}\right)^2\right) \cdot \mathcal{N}\left(z_t \mid a\mu_{t-1}, \sigma_z^2 + (a\sigma_{t-1})^2\right)}_{\text{use SRG-6 again}} \\
&\propto \mathcal{N}\left(z_t \mid \mu_t, \sigma_t^2\right)
\end{aligned}
$$

with

$$
\begin{aligned}
\rho_t^2 &= a^2\sigma_{t-1}^2 + \sigma_z^2 \quad \text{(predicted variance)} \\
K_t &= \frac{c\rho_t^2}{c^2\rho_t^2 + \sigma_x^2} \quad \text{(Kalman gain)} \\
\mu_t &= \underbrace{a\mu_{t-1}}_{\text{prior prediction}} + K_t \cdot \underbrace{\left(x_t - ca\mu_{t-1}\right)}_{\text{prediction error}} \quad \text{(posterior mean)} \\
\sigma_t^2 &= (1 - c \cdot K_t)\, \rho_t^2 \quad \text{(posterior variance)}
\end{aligned}
$$

- Kalman filtering consists of computing/updating these last four equations for each new observation (x_t). This is a very efficient recursive algorithm to estimate the state z_t from all observations (until t).
- It turns out that it's also possible to get an analytical result for $p(x_t|x^{t-1})$, which is the **model evidence** in a filtering context. See optional slides for details.

Multi-dimensional Kalman Filtering

- The Kalman filter equations can also be derived for multidimensional state-space models. In particular, for the model

$$
\begin{aligned}
z_t &= Az_{t-1} + \mathcal{N}(0, \Gamma) \\
x_t &= Cz_t + \mathcal{N}(0, \Sigma)
\end{aligned}
$$

  the Kalman filter update equations for the posterior $p(z_t|x^t) = \mathcal{N}(z_t \mid \mu_t, V_t)$ are given by (see Bishop, pg.639)

$$
\begin{aligned}
P_t &= AV_{t-1}A^T + \Gamma && \text{(predicted variance)} \\
K_t &= P_tC^T \cdot \left(CP_tC^T + \Sigma\right)^{-1} && \text{(Kalman gain)} \\
\mu_t &= A\mu_{t-1} + K_t \cdot \left(x_t - CA\mu_{t-1}\right) && \text{(posterior state mean)} \\
V_t &= (I - K_tC)\, P_t && \text{(posterior state variance)}
\end{aligned}
$$

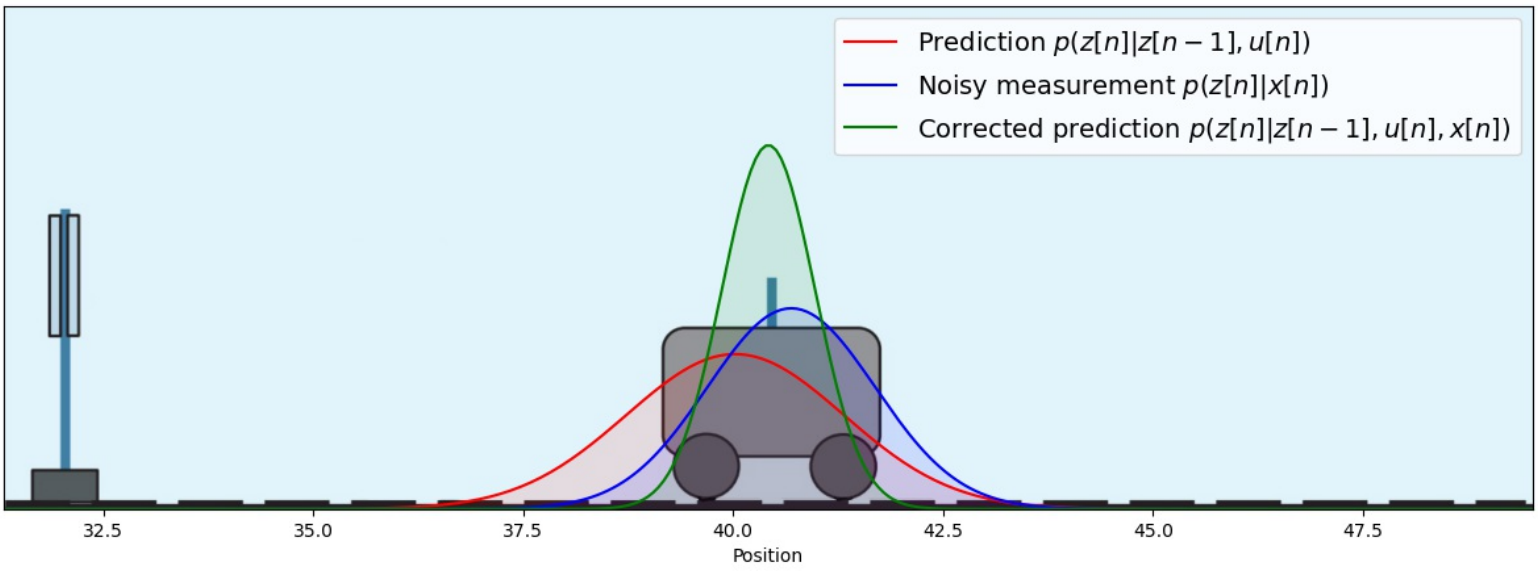
Code Example: Kalman Filtering and the Cart Position Tracking Example Revisited

- We can now solve the cart tracking problem of the introductory example by implementing the Kalman filter.

In [1]:
```
using Rocket, ReactiveMP, GraphPPL, LinearAlgebra, PyPlot
include("scripts/cart_tracking_helpers.jl")

# Specify the model parameters
Δt = 1.0                    # assume the time steps to be equal in size
A = [1.0 Δt;
     0.0 1.0]
b = [0.5*Δt^2; Δt]
Σz = convert(Matrix,Diagonal([0.2*Δt; 0.1*Δt])) # process noise covariance
Σx = convert(Matrix,Diagonal([1.0; 2.0]))      # observation noise covariance;

# Generate noisy observations
n = 10                # perform 10 timesteps
z_start = [10.0; 2.0] # initial state
u = 0.2 * ones(n)     # constant input u
noisy_x = generateNoisyMeasurements(z_start, u, A, b, Σz, Σx);

m_z = noisy_x[1]                                  # initial predictive mean
V_z = A * (1e8*Diagonal(I,2) * A') + Σz           # initial predictive covariance

for t = 2:n
```

```
        global m_z, V_z, m_pred_z, V_pred_z
        #predict
        m_pred_z = A * m_z + b * u[t]                    # predictive mean
        V_pred_z = A * V_z * A' + Σz                     # predictive covariance
        #update
        gain = V_pred_z * inv(V_pred_z + Σx)             # Kalman gain
        m_z = m_pred_z + gain * (noisy_x[t] - m_pred_z) # posterior mean update
        V_z = (Diagonal(I,2)-gain)*V_pred_z              # posterior covariance update
    end
    println("Prediction: ",MvNormalMeanCovariance(m_pred_z,V_pred_z))
    println("Measurement: ", MvNormalMeanCovariance(noisy_x[n],Σx))
    println("Posterior: ", MvNormalMeanCovariance(m_z,V_z))
    plotCartPrediction(m_pred_z[1], V_pred_z[1], m_z[1], V_z[1], noisy_x[n][1], Σx[1][1]);
```

```
Prediction: MvNormalMeanCovariance(
μ: [40.02419016879109, 4.027163620410047]
Σ: [1.2958787328575079 0.3921572953097835; 0.3921572953130242 0.3415636711134632]
)

Measurement: MvNormalMeanCovariance(
μ: [40.69330578387033, 4.399420397305405]
Σ: [1.0 0.0; 0.0 2.0]
)

Posterior: MvNormalMeanCovariance(
μ: [40.42123562376644, 4.172595492179203]
Σ: [0.5516100293973586 0.15018972175285758; 0.15018972175409862 0.24143326063188655]
)
```



The Cart Tracking Problem Revisited: Inference by Message Passing

- Let's solve the cart tracking problem by sum-product message passing in a factor graph like the one depicted above. All we have to do is create factor nodes for the state-transition model $p(z_t|z_{t-1})$ and the observation model $p(x_t|z_t)$. Then we let ReactiveMP execute the message passing schedule.

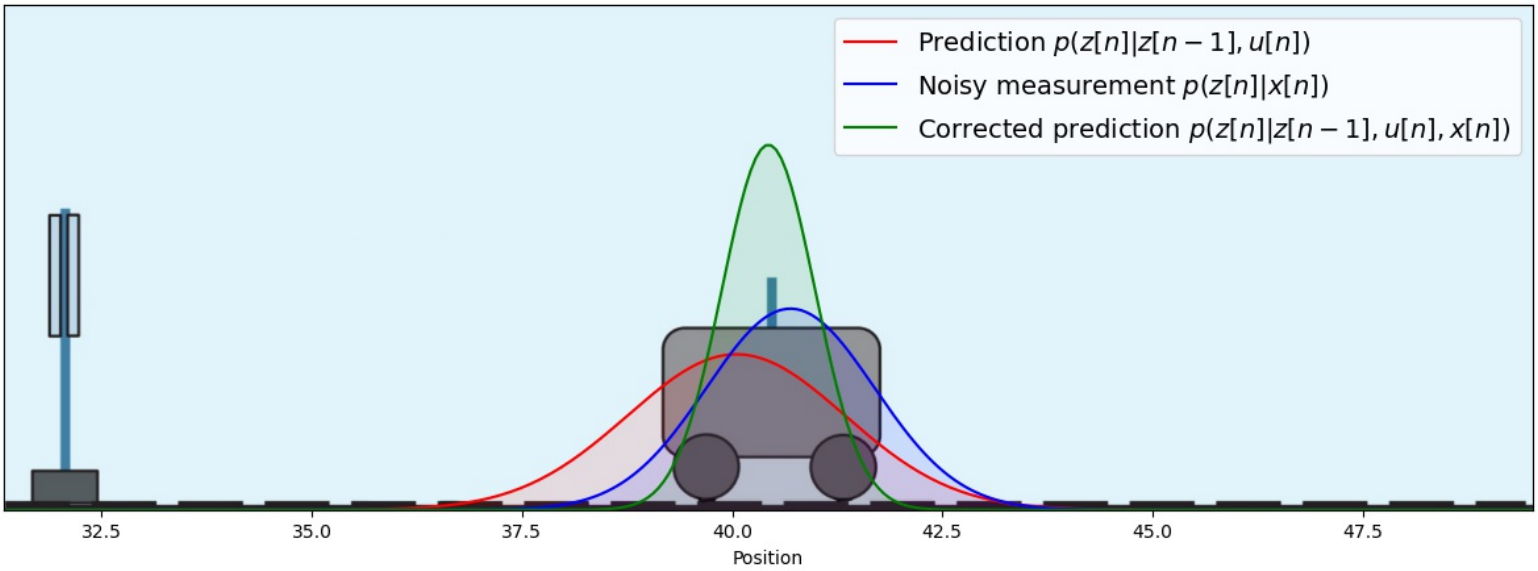
In [2]:
```
@model function cart_tracking(n, A, b, Σz, Σx, z_prev_m_0, z_prev_v_0, u)

    # We create constvar references for better efficiency
    cA = constvar(A)
    cB = constvar(b)
    cΣz = constvar(Σz)
    cΣx = constvar(Σx)

    znodes = Vector{FactorNode}(undef, n)
    # `z` is a sequence of hidden states
    z = randomvar(n)
    # `x` is a sequence of "clamped" observations
    x = datavar(Vector{Float64}, n)

    z_prior ~ MvNormalMeanCovariance(z_prev_m_0, z_prev_v_0)

    z_prev = z_prior
```

```
        for i in 1:n
            znodes[i],z[i] ~ MvNormalMeanCovariance(cA * z_prev + cB*u[i], cΣz)
            x[i] ~ MvNormalMeanCovariance(z[i], cΣx)
            z_prev = z[i]
        end

        return z, x, znodes
    end
```

Now that we've built the model, we can perform Kalman filtering by inserting measurement data into the model and performing message passing.

In [3]:
```
z_prev_m_0 = noisy_x[1]
z_prev_v_0 = A * (1e8*Diagonal(I,2) * A') + Σz ;
result = inference(model=Model(cart_tracking, n, A,b, Σz, Σx, z_prev_m_0, z_prev_v_0,u), data=(x=noisy_x,), free_ener
μz_posterior, Σz_posterior = mean.(result.posteriors[:z])[end], cov.(result.posteriors[:z])[end];
prediction_z_1 =  messageout(getinterface(result.returnval[end][end], :out))
prediction = ReactiveMP.materialize!(Rocket.getrecent(prediction_z_1));
println("Prediction: ",MvNormalMeanCovariance(mean(prediction), cov(prediction)))
println("Measurement: ", MvNormalMeanCovariance(noisy_x[n], Σx))
println("Posterior: ", MvNormalMeanCovariance(μz_posterior, Σz_posterior))
plotCartPrediction(mean(prediction)[1], cov(prediction)[1], μz_posterior[1], Σz_posterior[1], noisy_x[n][1], Σx[1][1]
```

```
Prediction: MvNormalMeanCovariance(
μ: [40.04920228769079, 4.040015184825758]
Σ: [1.2934227334046857 0.3916229823498387; 0.3916229823498387 0.3414332606222485]
)

Measurement: MvNormalMeanCovariance(
μ: [40.69330578387033, 4.399420397305405]
Σ: [1.0 0.0; 0.0 2.0]
)

Posterior: MvNormalMeanCovariance(
μ: [40.43118237832145, 4.1801088610998764]
Σ: [0.551150997075792 0.1501469959500068; 0.1501469959500068 0.24141815274489328]
)
```



- Note that both the analytical Kalman filtering solution and the message passing solution lead to the same results. The advantage of message passing-based inference with ReactiveMP is that we did not need to derive any inference equations. ReactiveMP took care of all that.

Recap Dynamical Models

- Dynamical systems do not obey the sample-by-sample independence assumption, but still can be specified, and state and parameter estimation equations can be solved by similar tools as for static models.
- Two of the more famous and powerful models with latent states include the hidden Markov model (with discrete states) and the Linear Gaussian dynamical system (with continuous states).
- For the LGDS, the Kalman filter is a well-known recursive state estimation procedure. The Kalman filter can be derived through Bayesian update rules on

Gaussian distributions.

- If anything changes in the model, e.g., the state noise is not Gaussian, then you have to re-derive the inference equations again from scratch and it may not lead to an analytically pleasing answer.

- ⇒ Generally, we will want to automate the inference processes. As we discussed, message passing in a factor graph is a visually appealing method to automate inference processes. We showed how Kalman filtering emerged naturally by automated message passing.

OPTIONAL SLIDES

Proof of Kalman filtering equations including evidence updating

- Now let's proof the Kalman filtering equations including evidence updating by probabilistic calculus:

$$\begin{aligned}
\overbrace{p(z_t \mid x^t)}^{\text{posterior}} &= p(z_t \mid x_t, x^{t-1}) \\
&= \frac{p(x_t, z_t \mid x^{t-1})}{p(x_t \mid x^{t-1})} \\
&= \frac{p(x_t \mid z_t)\, p(z_t \mid x^{t-1})}{p(x_t \mid x^{t-1})} \\
&= \frac{p(x_t \mid z_t) \sum_{z_{t-1}} p(z_t, z_{t-1} \mid x^{t-1})}{p(x_t \mid x^{t-1})} \\
&= \frac{p(x_t \mid z_t) \sum_{z_{t-1}} p(z_t \mid z_{t-1})\, p(z_{t-1} \mid x^{t-1})}{p(x_t \mid x^{t-1})} \\
&= \frac{\overbrace{\mathcal{N}(x_t \mid c z_t, \sigma_x^2)}^{\text{likelihood}} \sum_{z_{t-1}} \overbrace{\mathcal{N}(z_t \mid a z_{t-1}, \sigma_z^2)}^{\text{state transition}} \overbrace{\mathcal{N}(z_{t-1} \mid \mu_{t-1}, \sigma_{t-1}^2)}^{\text{prior}}}{\underbrace{p(x_t \mid x^{t-1})}_{\text{evidence}}}
\end{aligned}$$

- The posterior $p(z_t \mid x^t)$ is proportional to the numerator, which by making use of the renormalization equality

$$\mathcal{N}(x \mid cz, \sigma^2) = \frac{1}{c} \cdot \mathcal{N}\left(z \mid \frac{x}{c}, \left(\frac{\sigma}{c}\right)^2\right), \qquad \text{(renormalization)}$$

can be evaluated with Gaussian multiplication rules:

$$\begin{aligned}
&\mathcal{N}(x_t \mid c z_t, \sigma_x^2) \sum_{z_{t-1}} \underbrace{\mathcal{N}(z_t \mid a z_{t-1}, \sigma_z^2)}_{\text{use renormalization}} \mathcal{N}(z_{t-1} \mid \mu_{t-1}, \sigma_{t-1}^2) \\
&= \mathcal{N}(x_t \mid c z_t, \sigma_x^2) \sum_{z_{t-1}} \underbrace{\frac{1}{a} \mathcal{N}\left(z_{t-1} \mid \frac{z_t}{a}, \left(\frac{\sigma_z}{a}\right)^2\right) \mathcal{N}(z_{t-1} \mid \mu_{t-1}, \sigma_{t-1}^2)}_{\text{use Gaussian multiplication formula SRG-6}} \\
&= \frac{1}{a} \mathcal{N}(x_t \mid c z_t, \sigma_x^2) \sum_{z_{t-1}} \underbrace{\mathcal{N}\left(\mu_{t-1} \mid \frac{z_t}{a}, \left(\frac{\sigma_z}{a}\right)^2 + \sigma_{t-1}^2\right)}_{\text{not a function of } z_{t-1}} \underbrace{\mathcal{N}(z_{t-1} \mid \cdot, \cdot)}_{\text{sums to 1}} \\
&= \frac{1}{a} \underbrace{\mathcal{N}(x_t \mid c z_t, \sigma_x^2)}_{\text{use renormalization rule}} \underbrace{\mathcal{N}\left(\mu_{t-1} \mid \frac{z_t}{a}, \left(\frac{\sigma_z}{a}\right)^2 + \sigma_{t-1}^2\right)}_{\text{use renormalization rule}} \\
&= \underbrace{\frac{1}{c} \mathcal{N}\left(z_t \mid \frac{x_t}{c}, \left(\frac{\sigma_x}{c}\right)^2\right) \mathcal{N}\left(z_t \mid a\mu_{t-1}, \sigma_z^2 + (a\sigma_{t-1})^2\right)}_{\text{use SRG-6 again}} \\
&= \underbrace{\frac{1}{c} \mathcal{N}\left(\frac{x_t}{c} \mid a\mu_{t-1}, \left(\frac{\sigma_x}{c}\right)^2 + \sigma_z^2 + (a\sigma_{t-1})^2\right)}_{\text{use renormalization}} \mathcal{N}\left(z_t \mid \mu_t, \sigma_t^2\right) \\
&= \underbrace{\mathcal{N}\left(x_t \mid c a \mu_{t-1}, \sigma_x^2 + c^2(\sigma_z^2 + a^2 \sigma_{t-1}^2)\right)}_{\text{evidence } p(x_t \mid x^{t-1})} \underbrace{\mathcal{N}\left(z_t \mid \mu_t, \sigma_t^2\right)}_{\text{posterior } p(z_t \mid x^t)}
\end{aligned}$$

with

$$\begin{aligned}
\rho_t^2 &= a^2 \sigma_{t-1}^2 + \sigma_z^2 && \text{(predicted variance)} \\
K_t &= \frac{c\rho_t^2}{c^2 \rho_t^2 + \sigma_x^2} && \text{(Kalman gain)} \\
\mu_t &= \underbrace{a\mu_{t-1}}_{\text{prior prediction}} + K_t \cdot \underbrace{(x_t - c a \mu_{t-1})}_{\text{prediction error}} && \text{(posterior mean)} \\
\sigma_t^2 &= (1 - c \cdot K_t)\, \rho_t^2 && \text{(posterior variance)}
\end{aligned}$$

Intelligent Agents and Active Inference

Preliminaries

- Goal
  - Introduction to Active Inference and application to the design of synthetic intelligent agents

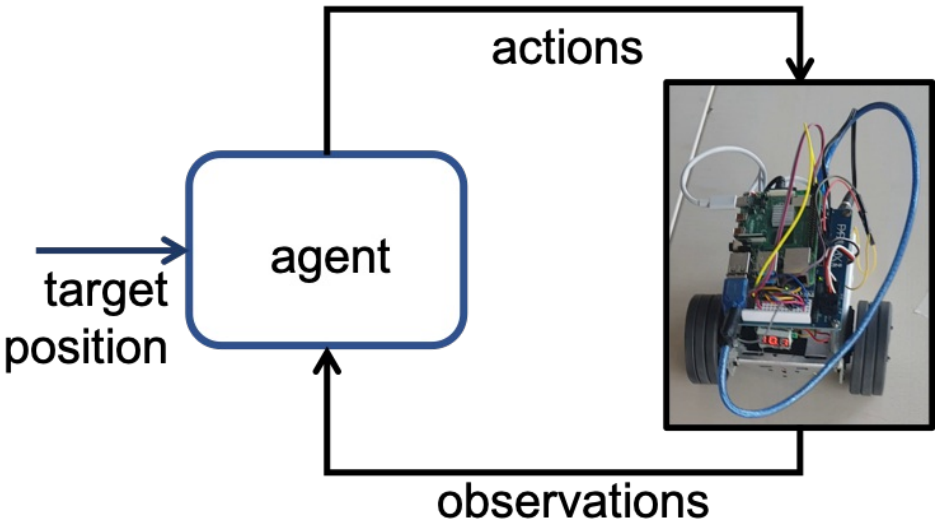
- Materials
  - Mandatory
    - These lecture notes
    - Karl Friston - 2016 - The Free Energy Principle (video)
  - Optional
    - Raviv (2018), The Genius Neuroscientist Who Might Hold the Key to True AI.
      - Interesting article on Karl Friston, who is a leading theoretical neuroscientist working on a theory that relates life and intelligent behavior to physics (and Free Energy minimization). (**highly recommended**)
    - Kirsch (2019), Theories of Intelligence: Active Inference
      - A nice tutural blog on active inference.
    - Van de Laar and De Vries (2019), Simulating Active Inference Processes by Message Passing
      - How to implement active inference by message passing in a Forney-style factor graph.

Agents

- In the previous lessons we assumed that a data set was given.
- In this lesson we consider *agents*. An agent is a system that *interacts* with its environment through both sensors and actuators.
- Crucially, by acting onto the environment, the agent is able to affect the data that it will sense in the future.
  - As an example, by changing the direction where I look, I can affect the (visual) data that will be sensed by my retina.
- With this definition of an agent, (biological) organisms are agents, and so are robots, self-driving cars, etc.
- In an engineering context, we are particularly interesting in agents that behave with a *purpose* (with a goal in mind), e.g., to drive a car or to design a speech recognition algorithm.
- In this lesson, we will describe how **goal-directed behavior** by biological (and synthetic) agents can also be interpreted as minimization of a free energy functional.

Illustrative Example: Steering a cart to a parking spot

- In this example, we consider a cart that can move in a 1D space.
- At each time step the cart can be steered a bit to the left or right by a controller (the "agent"). The agent's knowledge about the cart's process dynamics (equations of motion) are known up to some additive Gaussian process noise. The agent also makes noisy observations of the position and velocity of the cart.
- Your challenge is to design an agent that steers the car to the zero position. (The agent should be specified as a probabilistic model and the control signal should be formulated as a Bayesian inference task).



- Solution at the end of this lesson.

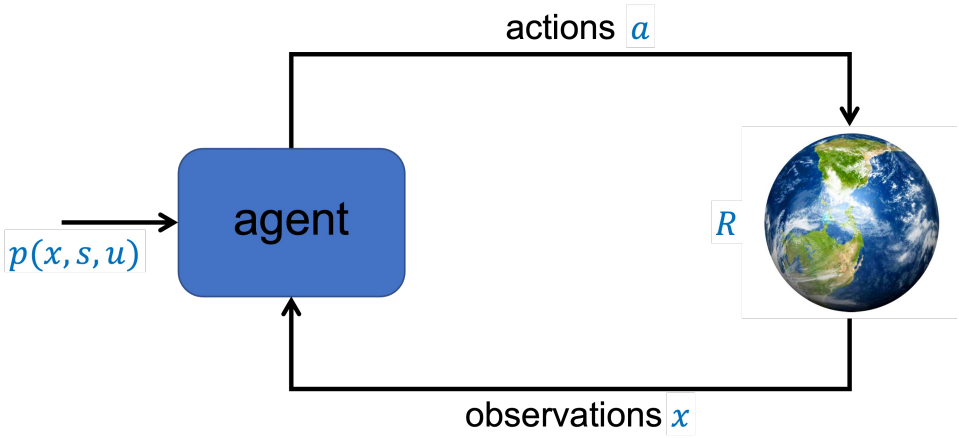
Karl Friston and the Free Energy Principle

- We begin with a motivating example that requires "intelligent" goal-directed decision making: assume that you are an owl and that you're hungry. What are you going to do?
- Have a look at Prof. Karl Friston's answer in this video segment on the cost function for intelligent behavior. (**Do watch the video!**)
- Friston argues that intelligent decision making (behavior, action making) by an agent requires *minimization of a functional of beliefs*.
- Friston further argues (later in the lecture and his papers) that this functional is a (variational) free energy (to be defined below), thus linking decision-making and acting to Bayesian inference.

- In fact, Friston's **Free Energy Principle** (FEP) claims that all biological self-organizing processes (including brain processes) can be described as Free Energy minimization in a probabilistic model.
    - This includes perception, learning, attention mechanisms, recall, acting and decision making, etc.
- Taking inspiration from FEP, if we want to develop synthetic "intelligent" agents, we have (only) two issues to consider:
    1. Specification of the FE functional.
    2. *How* to minimize the FE functional (often in real-time under situated conditions).
- Agents that follow the FEP are said to be involved in **Active Inference** (AIF). An AIF agent updates its states and parameters (and ultimately its model structure) solely by FE minimization, and selects its actions through (expected) FE minimization (to be explained below).

Execution of an AIF Agent

- Consider an AIF agent with observations (sensory states) x_t, latent internal states s_t and latent control states u_t for $t = 1, 2, \ldots$.



- The agent is embedded in an environment with "external states" $\tilde{s}_t$. The dynamics of the environment are driven by actions.
- Actions a_t are selected by the agent. Actions affect the environment and consequently affect future observations.
- In pseudo-code, an AIF agent executes the following algorithm:

    ACTIVE INFERENCE (AIF) AGENT ALGORITHM

    SPECIFY generative model $p(x, s, u)$
    ASSUME/SPECIFY environmental process R

    FORALL t DO

    1. $(x_t, \tilde{s}_t) = R(a_t, \tilde{s}_{t-1})$ % environment generates new observation
    2. $q(s_t) = \arg\min_q F[q]$ % update internal states (process observation)
    3. $q(u_{t+1}) = \arg\min_q H[q]$ % update control states (process observation)
    4. $\hat{u}_{t+1} \sim q(u_{t+1})$; $a_{t+1} = \hat{u}_{t+1}$ % sample next action and push to environment

    END

- In the above algorithm, $F[q]$ and $H[q]$ are appropriately defined Free Energy functionals, to be discussed below. Next, we discuss these steps in more details.

Specification of AIF Agent's model and Environmental Dynamics

- What should the agent's model p be modeling? This question was (already) answered by Conant and Ashby (1970) as the *good regulator theorem*: **every good regulator of a system must be a model of that system**.
- So the agent's model p will be a model that aims to explain how environmental causes (latent states) lead to observations, see the OPTIONAL SLIDE for more information.
- In this notebook, for illustrative purposes, we specify the **generative model** at time step t of an AIF agent as

$$p(x_t, s_t, u_t | s_{t-1}) = \underbrace{p(x_t|s_t)}_{\text{observations}} \cdot \underbrace{p(s_t|s_{t-1}, u_t)}_{\substack{\text{state} \\ \text{transition}}} \cdot \underbrace{p(u_t)}_{\substack{\text{action} \\ \text{prior}}}$$

- We also assume that the agent interacts with an environment, which we represent by a dynamic model R as

$$(x_t, \tilde{s}_t) = R(a_t, \tilde{s}_{t-1})$$

where a_t are *actions* , x_t are *outcomes* (the agent's observations) and $\tilde{s}_t$ holds the environmental latent *states*.

- The agent's knowledge about environmental process R is expressed by its generative model $p(x_t, s_t, u_t|s_{t-1})$.
- Note that R only needs to be specified for simulated environments. If we were to deploy the agent in a real-world environment, we would not need to specify R.
- Note that we distinguish between *control states* and *actions*. Control states u_t are latent variables in the agent's generative model. An action a_t is a realization of a control state as observed by the environment.
- Observations $x_t = \hat{x}_t$ are generated by the environment and observed by the agent. Vice versa, actions $a_t = \hat{u}_t$ are generated by the agent and observed by the environment.

State Updating in the AIF Agent

- After the agent makes a new observation $x_t = \hat{x}_t$, it will update beliefs over its latent variables. First the internal state variables s.
- Assume the following at time step t:
  - the state of the agent's model has already been updated to $q(s_{t-1}|\hat{x}_{1:t-1})$.
  - the agent has selected a new action $a_t = \hat{u}_t$.
  - the agent has recorded a new observation $x_t = \hat{x}_t$.
- The **state updating** task is to infer $q(s_t|x_{1:t})$, based on the previous estimate $q(s_{t-1}|x_{1:t-1})$, the new data $\{a_t = \hat{u}_t, x_t = \hat{x}_t\}$, and the agent's generative model.
- Technically, this is a Bayesian filtering task. In a real brain, this process is called **perception**.
- We specify the following FE functional

$$F[q] = \sum_{\tilde{s}_t} q(s_t|x_{1:t}) \log \frac{\overbrace{q(s_t|\hat{x}_{1:t})}^{\text{state posterior}}}{\underbrace{p(x_t|s_t)p(s_t|s_{t-1}, u_t)}_{\text{generative model w new data}} \underbrace{q(s_{t-1}|x_{1:t-1})}_{\text{state prior}}}$$

- The state updating task can be formulated as minimization of the above FE (see also AIF Algorithm, step 2):

$$q(s_t|x_{1:t}) = \arg\min_q F[q]$$

- In case the generative model is a *Linear Gaussian Dynamical System*, minimization of the FE can be solved analytically in closed-form and leads to the standard Kalman filter.
- In case these (linear Gaussian) conditions are not met, we can still minimize the FE by other means and arrive at some approximation of the Kalman filter, see for example Baltieri and Isomura (2021) for a Laplace approximation to variational Kalman filtering.
- Our toolbox `RxInfer.jl` specializes in automated execution of this minimization task.

Policy Updating in an AIF Agent

- Once the agent has updated its internal states, it will turn to inferring the next action.
- In order to select a **good** next action, we need to investigate and compare consequences of a *sequence* of future actions.
- A sequence of future actions $a = (a_{t+1}, a_{t+2}, \ldots, a_{t+T})$ is called a **policy**. Since relevant consequences are usually the result of an future action sequence rather than a single action, we will be interested in updating beliefs over policies.
- In order to assess the consequences of a selected policy, we will, as a function of that policy, run the generative model forward-in-time to make predictions about future observations $x_{t+1:T}$.
- Note that perception (state updating) preceeds policy updating. In order to accurately predict the future, the agent first needs to understand the current state of the world.
- Consider an AIF agent at time step t with (future) observations $x = (x_{t+1}, x_{t+2}, \ldots, x_{t+T})$, latent future internal states $s = (s_t, s_{t+1}, \ldots, s_{t+T})$, and latent future control variables $u = (u_{t+1}, u_{t+2}, \ldots, u_{t+T})$.
- From the agent's viewpoint, the evolution of these future variables are constrained by its generative model, rolled out into the future:

$$p(x, s, u) = \underbrace{q(s_t)}_{\substack{\text{current} \\ \text{state}}} \cdot \underbrace{\prod_{k=t+1}^{t+T} p(x_k|s_k) \cdot p(s_k|s_{k-1}, u_k) p(u_k)}_{\text{GM roll-out to future}}$$

- Consider the Free Energy functional for estimating posterior beliefs $q(s, u)$ over future states and control signals:

$$H[q] = \overbrace{\sum_{x,s} q(x|s)}^{\text{marginalize } x} \Big(\overbrace{\sum_u q(s,u) \log \frac{q(s,u)}{p(x,s,u)}}^{\text{variational Free Energy}} \Big)$$
$$= \sum_{x,s,u} q(x,s,u) \log \frac{q(s,u)}{p(x,s,u)}$$

- In principle, this is a regular FE functional, with one difference to previous versions: since future observations x have not yet occurred, $H[q]$ marginalizes not only over latent states s and policies u, but also over future observations x.
- We will update the beliefs over policies by minimization of Free Energy functional $H[q]$. In the optional slides below, we prove that the solution to this optimization task is given by (see AIF Algorithm, step 3, above)

$$q^*(u) = \arg\min_q H[q]$$
$$\propto p(u) \exp(-G(u)),$$

  where the factor $p(u)$ is a prior over admissible policies, and the factor $\exp(-G(u))$ updates the prior with information about future consequences of a selected policy u.
- The function

$$G(u) = \sum_{x,s} q(x,s|u) \log \frac{q(s|u)}{p(x,s|u)}$$

  is called the **Expected Free Energy** (EFE) for policy u.
- The FEP takes the following stance: if FE minimization is all that an agent does, then the only consistent and appropriate behavior for an agent is to select actions that minimize the **expected** Free Energy in the future (where expectation is taken over current beliefs about future observations).
- Note that, since $q^*(u) \propto p(u)\exp(-G(u))$, the probability $q^*(u)$ for selecting a policy u increases when EFE $G(u)$ gets smaller.
- Once the policy (control) variables have been updated, in simulated environments, it is common to assume that the next action a_{t+1} (an action is the *observed* control variable by the environment) gets selected in proportion to the probability of the related control variable (see AIF Agent Algorithm, step 4, above), i.e.,

$$\hat{u}_{t+1} \sim q(u_{t+1}) \quad \text{(select control realization by sampling)}$$
$$a_{t+1} = \hat{u}_{t+1} \quad \text{(transfer to environment)}$$

- Next, we analyze some properties of the EFE.

Active Inference Analysis: exploitation-exploration dilemma

- Consider the following decomposition of EFE:

$$\begin{aligned} G(u) &= \sum_{x,s} q(x,s|u) \log \frac{q(s|u)}{p(x,s|u)} \\ &= \sum_{x,s} q(x,s|u) \log \frac{1}{p(x)} + \sum_{x,s} q(x,s|u) \log \frac{q(s|u)}{p(s|x,u)} \frac{q(s|x)}{q(s|x)} \\ &= \sum_{x} q(x|u) \log \frac{1}{p(x)} + \sum_{x,s} q(x,s|u) \log \frac{q(s|u)}{q(s|x)} + \underbrace{\sum_{x,s} q(x,s|u) \log \frac{q(s|x)}{p(s|x,u)}}_{E[D_{\mathrm{KL}}[q(s|x), p(s|x,u)]] \geq 0} \\ &\geq \underbrace{\sum_{x} q(x|u) \log \frac{1}{p(x)}}_{\substack{\text{goal-seeking behavior} \\ \text{(exploitation)}}} - \underbrace{\sum_{x,s} q(x,s|u) \log \frac{q(s|x)}{q(s|u)}}_{\substack{\text{information-seeking behavior} \\ \text{(exploration)}}} \end{aligned}$$

- Apparently, minimization of EFE leads to selection of policies that balances the following two imperatives:
  1. minimization of the first term of $G(u)$, i.e. minimizing $\sum_x q(x|u) \log \frac{1}{p(x)}$, leads to policies ($u$) that align the inferred observations $q(x|u)$ with the prior on observations $p(x)$. We are in control to choose any prior $p(x)$ and usually we choose a prior that aligns with desired (goal) observations. Hence, policies with low EFE leads to **goal-seeking behavior** (a.k.a. pragmatic behavior or exploitation). In the OPTIONAL SLIDES, we derive an alternative (perhaps clearer) expression to support this interpretation].
  2. minimization of $G(u)$ maximizes the second term

$$\sum_{x,s} q(x,s|u) \log \frac{q(s|x)}{q(s|u)} = \sum_{x,s} q(x,s|u) \log \frac{q(s|x)}{q(s|u)} \frac{q(x|u)}{q(x|u)}$$
$$= \underbrace{\sum_{x,s} q(x,s|u) \log \frac{q(x,s|u)}{q(x|u)q(s|u)}}_{\text{(conditional) mutual information } I[x,s|u]}$$

     which is the (conditional) **mutual information** between (posteriors on) observations and states, for a given policy u. Thus, maximizing this term leads to actions that maximize statistical dependency between observations and states. In other words, a policy with low EFE also leads to **information-seeking behavior** (a.k.a. epistemic behavior or exploration).
- (The third term $\sum_{x,s} q(x,s|u) \log \frac{q(s|x)}{p(s|x)}$ is an (expected) KL divergence between posterior and prior on the states. This can be interpreted as a complexity/regularization term and $G(u)$ minimization will drive this term to zero.)

- Seeking actions that balance goal-seeking behavior (exploitation) and information-seeking behavior (exploration) is a fundamental problem in the Reinforcement Learning literature.
- **Active Inference solves the exploration-exploitation dilemma**. Both objectives are served by EFE minimization without need for any tuning parameters.

AIF Agents learn both the Problem and Solution

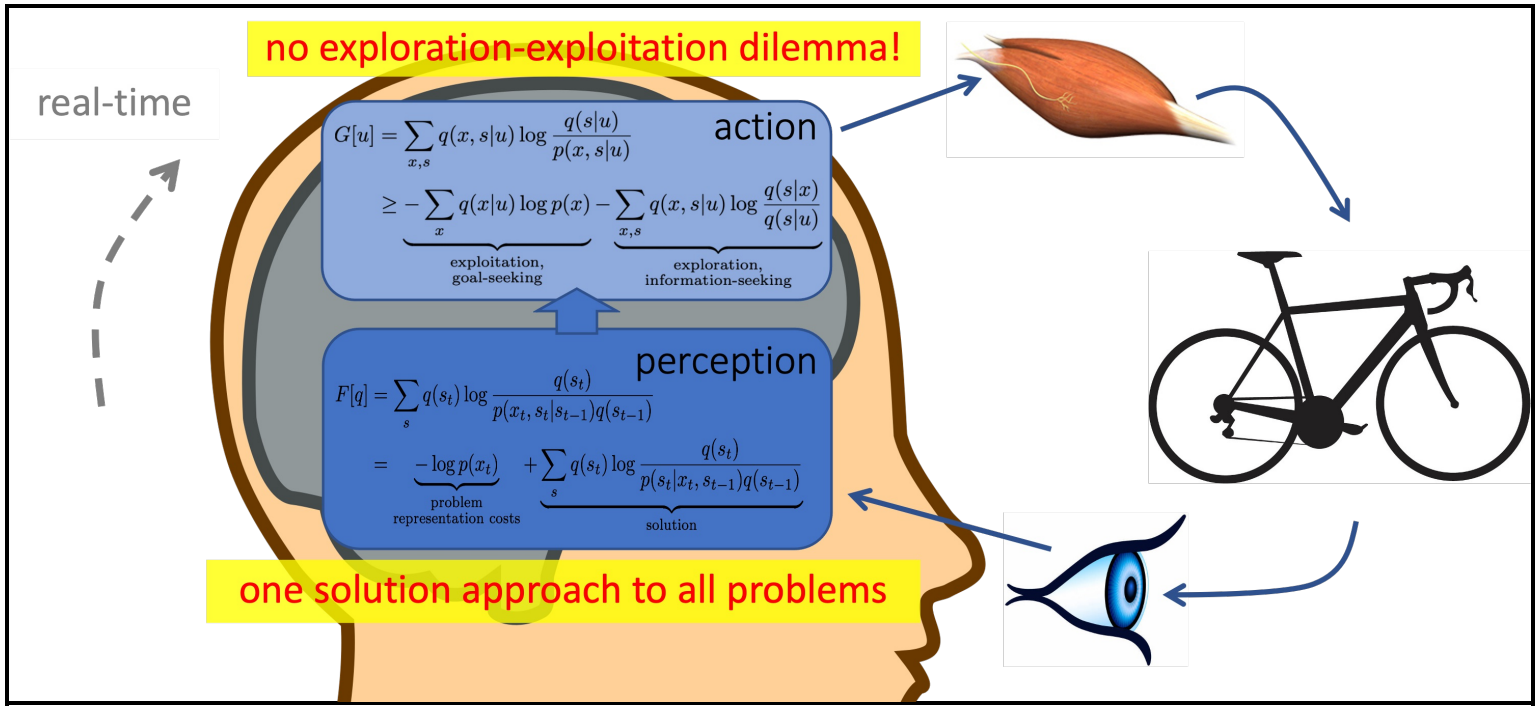
- We highlight another great feature of FE minimizing agents. Consider an AIF agent (m) with generative model $p(x, s, u|m)$.
- Consider the Divergence-Evidence decomposition of the FE again:

$$\begin{aligned} F[q] &= \sum_{s,u} q(s,u) \log \frac{q(s,u)}{p(x,s,u|m)} \\ &= \underbrace{-\log p(x|u,m)}_{\substack{\text{problem} \\ \text{representation costs}}} + \underbrace{\sum_{s,u} q(s,u) \log \frac{q(s,u)}{p(s,u|x,m)}}_{\text{solution costs}} \end{aligned}$$

- The first term, $-\log p(x|u,m)$, is the (negative log-) evidence for model m, given data x that has been recorded under policy u.
- Minimization of FE maximizes the evidence for the given model. The model captures the **problem representation**. A model with high evidence predicts the data well and therefore "understands the world".
- The second term scores the cost of inference. In almost all cases, the solution to a problem can be phrased as an inference task on the generative model. Hence, the second term **scores the accuracy of the inferred solution**, for the given model.
- FE minimization optimizes a balanced trade-off between a good-enough problem representation and a good-enough solution proposal for that model. Since FE comprises both a cost for solution *and* problem representation, it is a neutral criterion that applies across a very wide set of problems.
- A good solution to the wrong problem is not good enough. A poor solution to a great problem statement is not sufficient either. In order to solve a problem well, we need both to represent the problem correctly (high model evidence) and we need to solve it well (low inference costs).

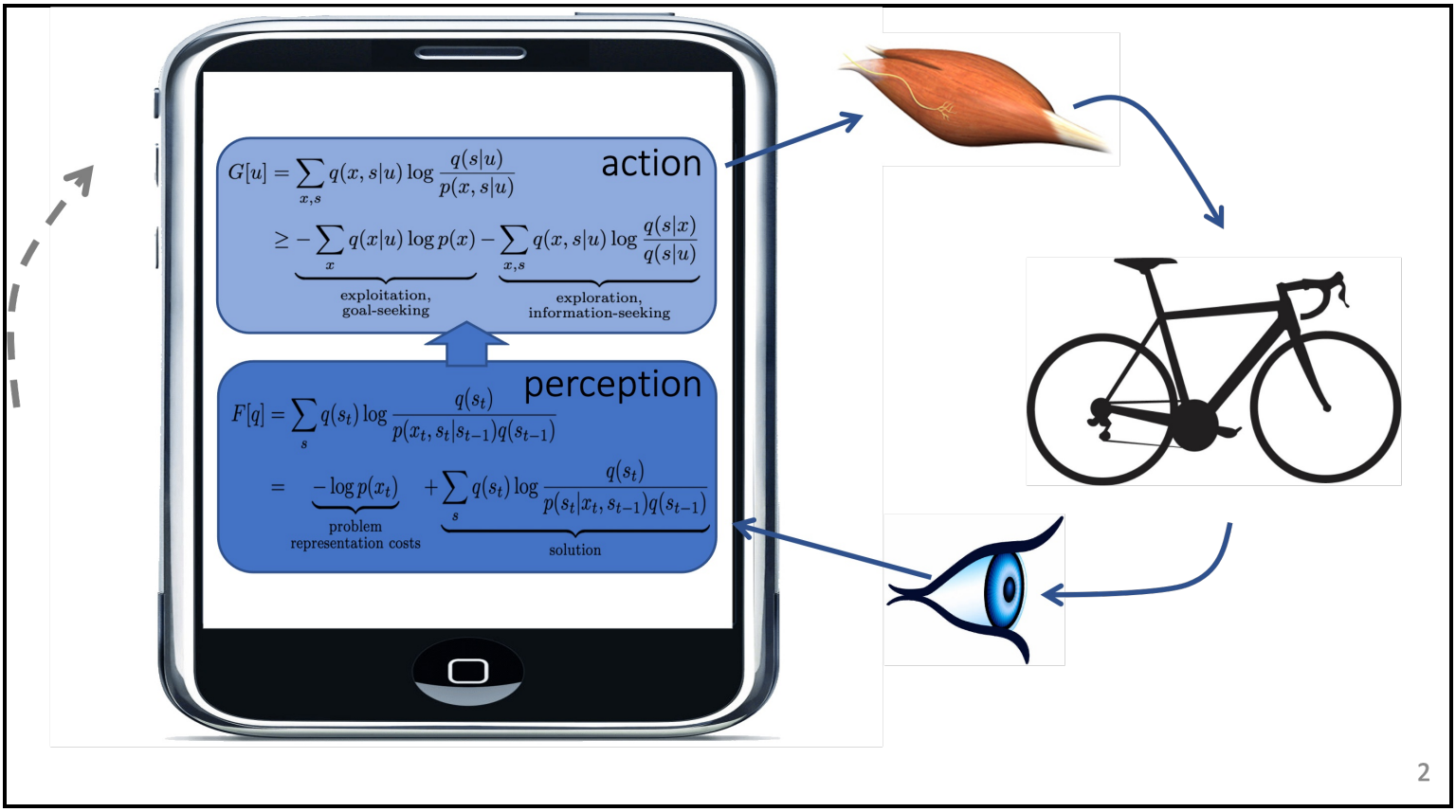
The Brain's Action-Perception Loop by FE Minimization

- The above derivations are not trivial, but we have just shown that FE minimizing agents accomplish variational Bayesian perception (a la Kalman filtering), and a balanced exploration-exploitation trade-off for policy selection.
- Moreover, the FE by itself serves as a proper objective across a very wide range of problems, since it scores both the cost of the problem statement and the cost of inferring the solution.
- The current FEP theory claims that minimization of FE (and EFE) is all that brains do, i.e., FE minimization leads to perception, policy selection, learning, structure adaptation, attention, learning of problems and solutions, etc.



The Engineering Challenge: Synthetic AIF Agents

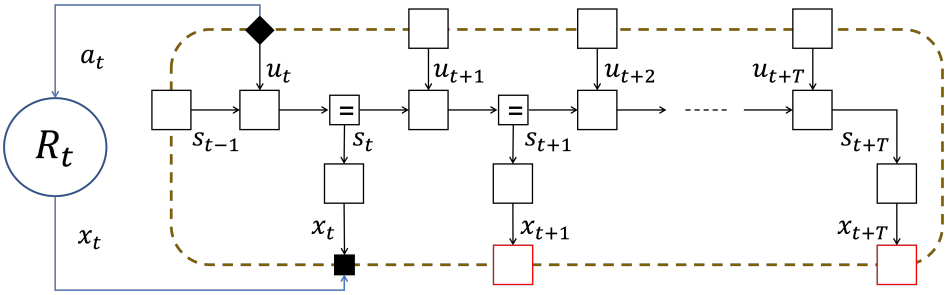
- So we have here an AI framework (minimization of FE and associated EFE) that
  - leads to optimal (Bayesian) information processing, including balancing accuracy vs complexity.
  - leads to balanced and continual learning of both problem representation and solution proposal
  - actively selects data in-the-field under situated conditions (no dependency on large data base)
  - pursues a optimal trade-off between exploration (information-seeking) and exploitation (goal-seeking) behavior
  - needs no external tuning parameters (such as step sizes, thresholds, etc.)
  - has a strong foundations in how nature/life self-organizes
- Clearly, the FEP, and synthetic AIF agents as a realization of FEP, comprise a very attractive framework for all things relating to AI and AI agents.
- A current big AI challenge is to design synthetic AIF agents based solely on FE/EFE minimization.



- Executing a synthetic AIF agent often poses a large computational problem because of the following reasons:
  1. For interesting problems (e.g. speech recognition, scene analysis), generative models may contain thousands of latent variables.
  2. The FE function is a time-varying function, since it is also a function of observable variables.
  3. An AIF agent must execute inference in real-time if it is engaged and embedded in a real world environment.
- So, in practice, executing a synthetic AIF agent may lead to a **task of minimizing a time-varying FE function of thousands of variables in real-time**!!

Factor Graph Approach to Modeling of an Active Inference Agent

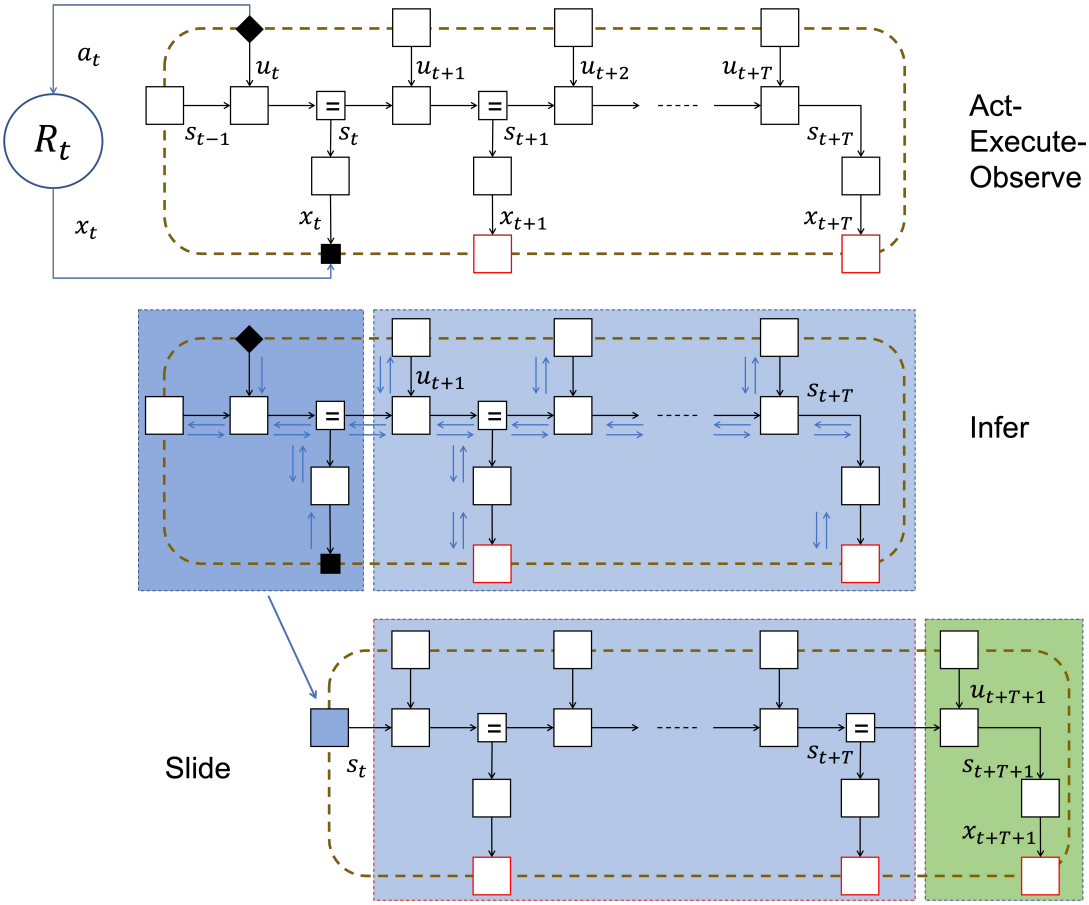
- How to specify and execute a synthetic AIF agent is an active area of research.
- There is no definitive solution approach to AIF agent modeling yet; we (BIASlab) think that (reactive) message passing in a factor graph representation provides a promising path.
- After selecting an action a_t and making an observation x_t, the FFG for the rolled-out generative model is given by the following FFG:



- The open red nodes for $p(x_{t+k})$ specify **desired future observations**, whereas the open black boxes for $p(s_k|s_{k-1}, u_k)$ and $p(x_k|s_k)$ reflect the agent's beliefs about how the world actually evolves (ie, the **veridical model**).
- The (brown) dashed box is the agent's Markov blanket. Given the states on the Markov blanket, the internal states of the agent are independent of the state of the world.

How to minimize FE: Online Active Inference

- Online active inference proceeds by iteratively executing three stages:
  1. act-execute-observe
  2. update the latent variables and select an action
  3. slide forward



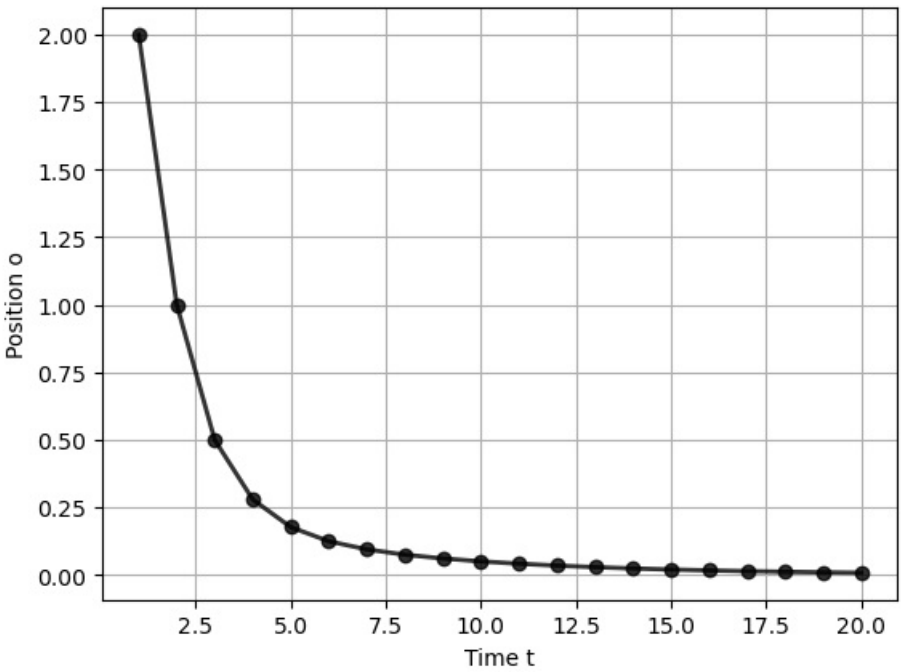
The Cart Parking Problem Revisited

Here we solve the cart parking problem as stated at the beginning of this lesson. We first specify a generative model for the agent's environment (which is the observed noisy position of the cart) and then constrain future observations by a prior distribution that is located on the target parking spot. Next, we schedule a message passing-based inference algorithm for the next action. This is followed by executing the *Act-execute-observe --> infer --> slide* procedure to infer a sequence of consecutive actions. Finally, the position of the cart over time is plotted. Note that the cart converges to the target spot.

In [1]:
```
using Pkg;Pkg.activate("probprog/workspace");Pkg.instantiate()
IJulia.clear_output();
```

In [2]:
```
using PyPlot, LinearAlgebra, ForneyLab
```

```
# Load helper functions. Feel free to explore these
include("ai_agent/environment_1d.jl")
include("ai_agent/helpers_1d.jl")
include("ai_agent/agent_1d.jl")

# Internal model perameters
gamma   = 100.0 # Transition precision
phi     = 10.0 # Observation precision
upsilon = 1.0 # Control prior variance
sigma   = 1.0 # Goal prior variance

T = 10 # Lookahead

# Build internal model
fg = FactorGraph()

o = Vector{Variable}(undef, T) # Observed states
s = Vector{Variable}(undef, T) # internal states
u = Vector{Variable}(undef, T) # Control states

@RV s_t_min ~ GaussianMeanVariance(placeholder(:m_s_t_min),
                                   placeholder(:v_s_t_min)) # Prior  state
u_t = placeholder(:u_t)
@RV u[1] ~ GaussianMeanVariance(u_t, tiny)
@RV s[1] ~ GaussianMeanPrecision(s_t_min + u[1], gamma)
@RV o[1] ~ GaussianMeanPrecision(s[1], phi)
placeholder(o[1], :o_t)

s_k_min = s[1]
for k=2:T
    @RV u[k] ~ GaussianMeanVariance(0.0, upsilon) # Control prior
    @RV s[k] ~ GaussianMeanPrecision(s_k_min + u[k], gamma) # State transition model
    @RV o[k] ~ GaussianMeanPrecision(s[k], phi) # Observation model
    GaussianMeanVariance(o[k],
                         placeholder(:m_o, var_id=:m_o_*k, index=k-1),
                         placeholder(:v_o, var_id=:v_o_*k, index=k-1)) # Goal prior
    s_k_min = s[k]
end

# Schedule message passing algorithm
algo = messagePassingAlgorithm(u[2]) # Infer internal states
source_code = algorithmSourceCode(algo)
eval(Meta.parse(source_code)) # Loads the step!() function for inference

s_0 = 2.0 # Initial State

N = 20 # Total simulation time

(execute, observe)  = initializeWorld() # Let there be a world
(infer, act, slide) = initializeAgent() # Let there be an agent

# Step through action-perception loop
u_hat = Vector{Float64}(undef, N) # Actions
o_hat = Vector{Float64}(undef, N) # Observations
for t=1:N
    u_hat[t] = act() # Evoke an action from the agent
               execute(u_hat[t]) # The action influences hidden external states
    o_hat[t] = observe() # Observe the current environmental outcome (update p)
               infer(u_hat[t], o_hat[t]) # Infer beliefs from current model state (update q)
               slide() # Prepare for next iteration
end

# Plot active inference results
plotTrajectory(u_hat, o_hat)
;
```

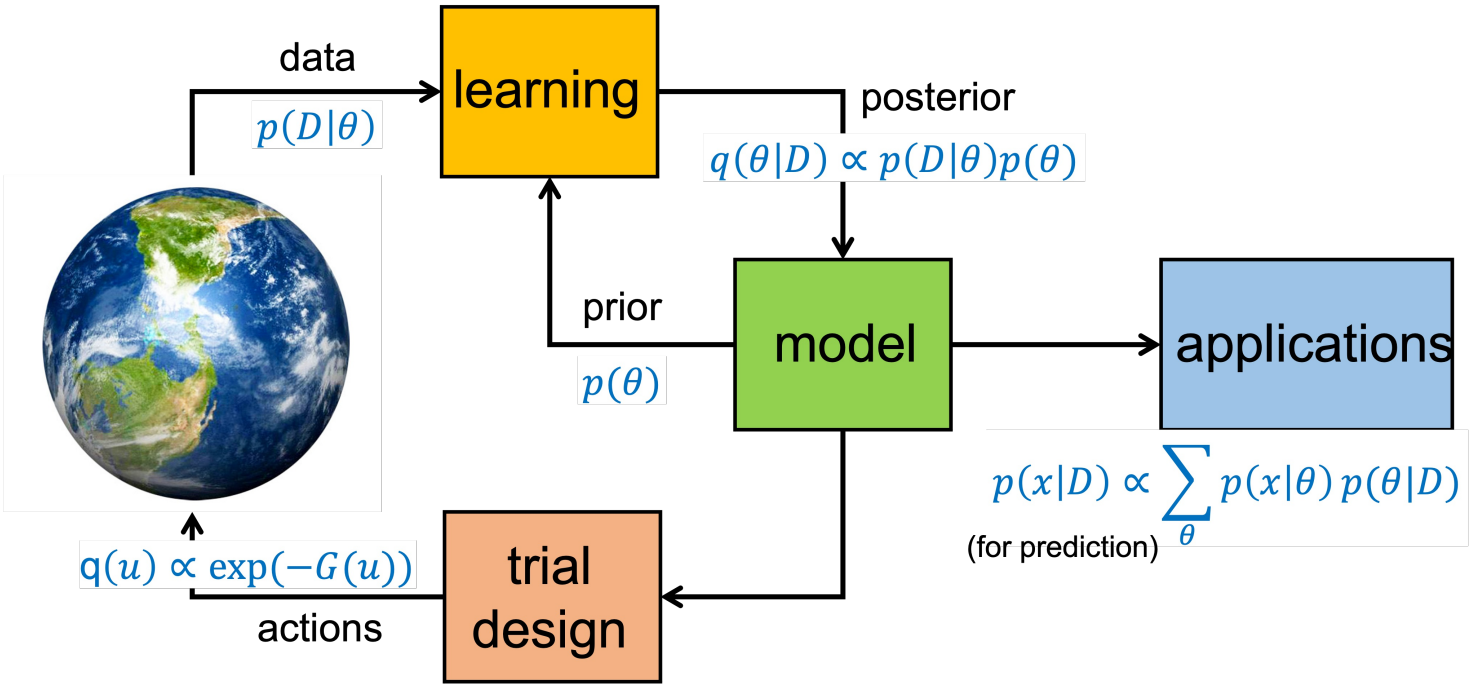


Video

- See this video to verfiy that the robot will be steered to the correct parking spot even after an "adversarial" intervention :). (This project was completed by Burak Ergul as part of this MSc graduation project).

Extensions and Comments

- If interested, here is a link to a more detailed version of the 1D parking problem.
- We also have a 2D version of this cart parking problem implemented on Raspberry Pi-based robot. (Credits for this implemention to Thijs van de Laar and Burak Ergul).
- Just to be sure, you don't need to memorize all FE/EFE decompositions nor are you expected to derive them on-the-spot. We present these decompositions only to provide insight into the multitude of forces that underlie FEM-based action selection.
- In a sense, the FEP is an umbrella for describing the mechanics and self-organization of intelligent behavior, in man and machines. Lots of sub-fields in AI, such as reinforcement learning, can be interpreted as a special case of active inference under the FEP, see e.g., Friston et al., 2009.
- Is EFE minimization really different from "regular" FE minimization? Not really, it appears that EFE minimization can be reformulated as a special case of FE minimization. In other words, FE minimization is still the only game in town.
- Active inference also completes the "scientific loop" picture. Under the FEP, experimental/trial design is driven by EFE minimization. Bayesian probability theory (and FEP) contains all the equations for running scientific inquiry.



- The big engineering challenge remains the computational load of AIF. The human brain consumes about 25 Watt and the neocortex only about 4 Watt (which is about the power consumption of a bicycle light). This is multiple orders of magnitude cheaper than what we can engineer on silicon for similar tasks.

Final Thoughts

- In the end, all the state inference, parameter estimation, etc., in this lecture series could have been implemented by FE minimization in an appropriately specified generative probabilistic model. However, the Free Energy Principle extends beyond state and parameter estimation. Driven by FE minimization, brains change their structure as well over time. In fact, the FEP extends beyond brains to a general theory for biological self-organization, e.g., Darwin's natural selection process may be interpreted as a FE minimization-driven model optimization process, and here's an article on FEP for predictive processing in plants. Moreover, Constrained-FE minimization (rephrased as the Principle of Maximum Relative Entropy) provides an elegant framework to derive most (if not all) physical laws, as Caticha exposes in his brilliant monograph on Entropic Physics. Indeed, the framework of FE minimization is known in the physics community as the very fundamental Principle of Least Action that governs the equations-of-motion in nature.
- So, the FEP is very fundamental and extends way beyond applications to machine learning. At our research lab at TU/e, we work on developing FEP-based intelligent agents that go out into the world and autonomously learn to accomplish a pre-determined task, such as learning-to-walk or learning-to-process-noisy-speech-signals. Free free to approach us if you want to know more about that effort.

OPTIONAL SLIDES

In an AIF Agent, Actions fulfill Desired Expectations about the Future

- In the derivations above, we decomposed the EFE into an upperbound on the sum of a goal-seeking and information-seeking term. Here, we derive an alternative (exact) decomposition that more clearly reveals the goal-seeking objective.
- We consider again the EFE and factorize the generative model $p(x,s|u) = p'(x)p(s|x,u)$ as a product of a **target prior** $p'(x)$ on observations and a **veridical** state model $p(s|x,u)$.
- Through the **target prior** $p'(x)$, the agent declares which observations it **wants** to observe in the future. (The prime is just to distinguish the semantics of a desired future from the model for the actual future).
- Through the **veridical** state model $p(s|x,u)$, the agent implicitly declares its beliefs about how the world will **actually** generate observations.
  - In particular, note that through the equality (by Bayes rule)

$$p(s|x,u) = \frac{p(x|s)p(s|u)}{p(x|u)} = \frac{p(x|s)p(s|u)}{\sum_s p(x|s)p(s|u)},$$

it follows that in practice the agent may specify $p(s|x,u)$ implicitly by explicitly specifying a state transition model $p(s|u)$ and observation model $p(x|s)$.

- Hence, an AIF agent holds both a model for its beliefs about how the world will actually evolve AND a model for its beliefs about how it desires the world to evolve!!

- To highlight the role of these two models in the EFE, consider the following alternative EFE decomposition:

$$
\begin{aligned}
G(u) &= \sum_{x,s} q(x,s|u) \log \frac{q(s|u)}{p'(x)p(s|x,u)} \\
&= \sum_{x,s} q(x,s|u) \log \frac{q(s|u)}{p'(x)} \frac{1}{p(s|x,u)} \\
&= \sum_{x,s} q(x,s|u) \log \frac{q(s|u)}{p'(x)} \frac{p(x|u)}{p(x|s)p(s|u)} \quad \text{(use Bayes)} \\
&= \sum_{x,s} q(x,s|u) \log \frac{q(s|u)}{p(x|s)p(s|u)} \frac{p(x|u)}{p'(x)} \\
&= \sum_{x,s} q(x,s|u) \log \frac{q(s|u)}{p(x|s)p(s|u)} + \sum_{x,s} q(x,s|u) \log \frac{p(x|u)}{p'(x)}
\end{aligned}
$$

$$
= \sum_{x,s} p(s|u)p(x|s) \log \frac{\cancel{p(s|u)}}{p(x|s)\,\cancel{p(s|u)}} + \sum_{x,s} p(s|u)p(x|s) \log \frac{p(x|u)}{p'(x)}
$$

$$
(\text{ assume } q(x,s|u) = p(x|s)p(s|u) \text{ })
$$

$$
\begin{aligned}
&= \sum_{s} p(s|u) \sum_{x} p(x|s) \log \frac{1}{p(x|s)} + \sum_{x} p(x|u) \log \frac{p(x|u)}{p'(x)} \\
&= \underbrace{E_{p(s|u)}\left[H[p(x|s)]\right]}_{\text{ambiguity}} + \underbrace{D_{\mathrm{KL}}\left[p(x|u), p'(x)\right]}_{\text{risk}}
\end{aligned}
$$

- In this derivation, we have assumed that we can use the generative model to make inferences in the "forward" direction. Hence, $q(s|u) = p(s|u)$ and $q(x|s) = p(x|s)$.

- The terms "ambiguity" and "risk" have their origin in utility theory for behavioral ecocomics. Minimization of EFE leads to minimizing both ambiguity and risk.

- Ambiguous (future) states are states that map to large uncertainties about (future) observations. We want to avoid those ambiguous states since it implies that the model is not capable to predict how the world evolves. Ambiguity can be resolved by selecting information-seeking (epistemic) actions.

- Minimization of the second term (risk) leads to choosing actions (u) that align **predicted** future observations (represented by $p(x|u)$) with **desired** future observations (represented by $p'(x)$). Agents minimize risk by selecting pragmatic (goal-seeking) actions.

- $\Rightarrow$ **Actions fulfill desired expectations about the future!**

- (return to related cell in main text).

Proof $q^*(u) = \arg\min_q H[q] \propto p(u)\exp(-G(u))$

- Consider the following decomposition:

$$
\begin{aligned}
H[q] &= \sum_{x,s,u} q(x,s,u) \log \frac{q(s,u)}{p(x,s,u)} \\
&= \sum_{x,s,u} q(x,s|u)q(u) \log \frac{q(s|u)q(u)}{p(x,s|u)p(u)} \\
&= \sum_{u} q(u) \left(\sum_{x,s} q(x,s|u) \log \frac{q(s|u)q(u)}{p(x,s|u)p(u)} \right) \\
&= \sum_{u} q(u) \left(\log q(u) + \log \frac{1}{p(u)} + \underbrace{\sum_{x,s} q(x,s|u) \log \frac{q(s|u)}{p(x,s|u)}}_{G(u)} \right) \\
&= \sum_{u} q(u) \log \frac{q(u)}{p(u)\exp(-G(u))}
\end{aligned}
$$

- This is a KL-divergence. Minimization of $H[q]$ leads to the following posterior for the policy:

$$
\begin{aligned}
q^*(u) &= \arg\min_q H[q] \\
&= \frac{1}{Z} p(u)\exp(-G(u))
\end{aligned}
$$

- (click to return to linked cell in the main text.)

What Makes a Good Agent? The Good Regulator Theorem

- According to Friston, an ``intelligent'' agent like a brain minimizes a variational free energy functional, which, in general, is a functional of a probability distribution p and a variational posterior q.

- What should the agent's model p be modeling? This question was (already) answered by Conant and Ashby (1970) as the Good Regulator Theorem: **every good regulator of a system must be a model of that system**.

- A Quote from Conant and Ashby's paper (this statement was later finessed by Friston (2013)):

> "The theory has the interesting corollary that the living brain, insofar as it is successful and efficient as a regulator for survival, *must*

proceed, in learning, by the formation of a model (or models) of its environment."

Int. J. Systems Sci., 1970, vol. 1, No. 2, 89-97

EVERY GOOD REGULATOR OF A SYSTEM MUST BE A MODEL OF THAT SYSTEM[1]

Roger C. Conant

Department of Information Engineering, University of Illinois, Box 4348, Chicago, Illinois, 60680, U.S.A.

and W. Ross Ashby

Biological Computers Laboratory, University of Illinois, Urbana, Illinois 61801, U.S.A.[2]

[Received 3 June 1970]

The design of a complex regulator often includes the making of a model of the system to be regulated. The making of such a model has hitherto been regarded as optional, as merely one of many possible ways.

m this paper a theorem is presented which shows, under very broad conditions, that any regulator that is maximally both successful and simple *must* be isomorphic with the system being regulated. (The exact assumptions are given.) Making a model is thus necessary.

The theorem has the interesting corollary that the living brain, so far as it is to be successful and efficient as a regulator for survival, *must* proceed, in learning, by the formation of a model (or models) of its environment.

- (Return to related cell in main text).